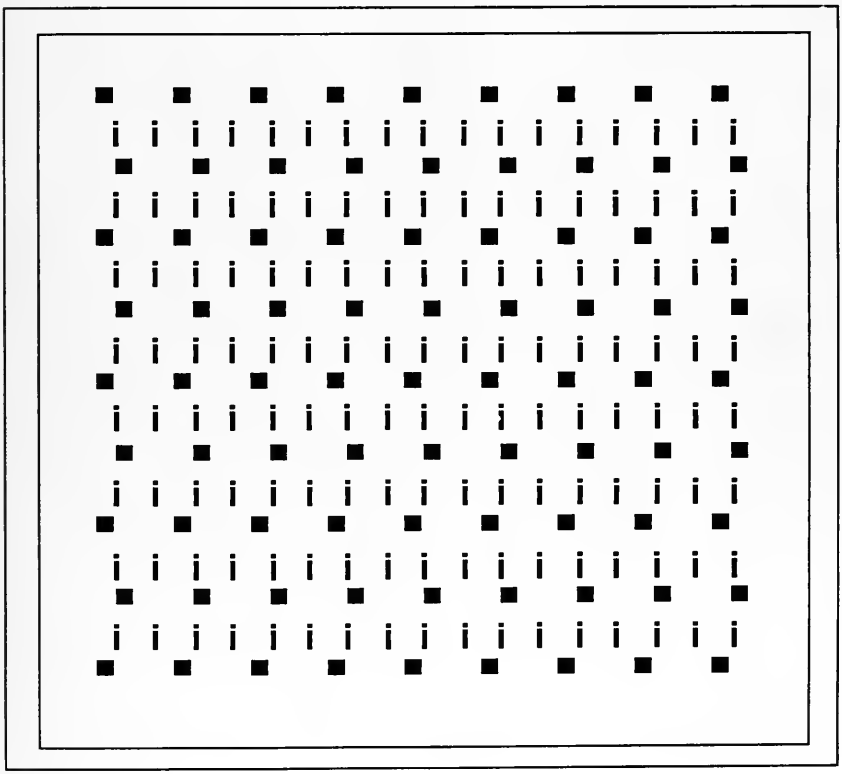
IASSIST

QUARTERLY

VOLUME 15 | Fall / Winter 1991 | NUMBER 3/4



Printed at UCLA

IASSIST QUARTERLY

The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors

The QUARTERLY is published four times per year. Articles and other information should be typewritten and double-spaced. Each page of the manuscript should be numbered. The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science 30(2):77-82, March 1979.* If the contribution is an announcement of a conference, training session, or the like, the text should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event. Book notices and reviews should not exceed two double-spaced pages. Deadlines for submitting articles are six weeks before publication. Manuscripts should be sent in duplicate to the Editor: Walter Piovesan, Research Data Library, W.A.C. Bennett Library, Simon Fraser University, Burnaby, B.C., V5A 1S6 CANADA. (604) 291-4349 E-Mail: USERDLIB@SFU.BITNET Book reviews should be submitted in duplicate to the Book Review Editor: Daniel Tsang, Main Library, University of California P.O. Box 19557, Irvine, California 92713 USA. (714) 856-4978 E-Mail: DTSANG@ORION.CF.UCI.EDU

Title: **Newsletter - International Association for Social Science Information Service and Technology**
ISSN - United States: 0739-1137

CONTENTS

Volume 15 Number 3/4 Fall/Winter1991

FEATURES

News

The Impact of Future Social and Technological Trends On the Dissemination of Census Bureau Information

by Donald L. Day [1]
School of Information Studies
Syracuse University

Abstract
This study examines social and technological trends that may impact the dissemination of U.S. census information via the Depository Library Program in the Year 2000 and beyond.

The study looks beyond currently emerging systems to examine a limited list of future issues in technology, regulation, funding, access, and user demand. It examines information dissemination in the broad, societal context, rather than concentrating narrowly upon the means of delivery. Its main objectives are to pinpoint key issues, to stimulate an appreciation of the inextricable nature of information in postindustrial society, and to recommend policies and directions for further research.

Introduction

Nature of the Topic
This study examines social and technological trends that may impact the dissemination of U.S. census information via the Depository Library Program in the Year 2000 and beyond.

Importance of the Topic
Establishing the social and technological context within which census information might be disseminated in the future would facilitate rational policy making regarding that dissemination. It also might save time and money, in that decisions about systems with long lead times could be made so as to implement the systems in a timely and efficient manner.

Scope and Objectives
The original impetus for the study was a concern about the sheer volume of paper products absorbed by the Depository Library Program (as high as 25 percent of all documents printed from the 1980 census), and an interest in how new technologies may make it possible to reduce costs while increasing the availability and variety of data (in particular, census data).

The study looks beyond currently emerging systems to examine a limited list of future issues in technology, regulation, funding, access, and user demand. It examines information dissemination in the broad, societal context, rather than concentrating narrowly upon the means of delivery. Its main objectives are to pinpoint key issues, to stimulate an appreciation of the inextricable nature of information in postindustrial society, and to recommend policies and directions for further research.

Point of View
This study was fielded under the presumption that government will be required to continue providing public access to federal information as part of its commitment to maintaining the informed citizenry that is central to participatory democracy.

Report Outline
Following a brief review of study methodology, this report presents an overview of future trends (the societal context), a discussion of study findings, and policy recommendations for key issues requiring public debate.

Study Method

Design
The current study was conceived as a qualitative, exploratory and descriptive effort to identify issues of concern. Elite interviewing was to be interspersed with a review of literature in the future studies field in an iterative process that would develop perspective and deepen focus and selectivity in data collection.

The interviews were in-depth but informal, as recommended by Marshall & Rossman (1989). They were conducted in subjects' normal work environment (a "natural setting") to ensure ease of discussion and immediate access to reference materials.

Procedure
Subjects were selected because of their expert knowledge of federal information dissemination policies and of developing technology. Interviews were conducted in two rounds, separated by approximately four weeks. The first set concentrated on preliminary data gathering and focused upon the Depository Library Program. The second set was guided by an outline of concerns (interrogative research questions) developed from the literature and from the earlier interviews (see below). Sessions were recorded to ensure the completeness of notes and to help recognize nuances that might have been overlooked at the time of initial data collection. Only one subject

was distracted by the presence of a taping machine; the others were largely oblivious to its presence.

The scheduling of multiple interview events separated by a review of the literature and by conceptual outlining worked very well in focusing the research and in identifying issues that were not apparent at the outset. An improvement in procedure that might be useful in other qualitative research on this or other topics would be submission of interim reports to key sources for critique.

This technique might help sources to feel involved in the research effort, tend to focus their comments during interviews, and encourage them to act as agents of the researcher in obtaining material useful to the study.

Key Research Questions

The key research questions drafted from an analysis of the literature and during the interviewing process were as follows.

1. What will be the leading edge information technologies in the first decade of the next century?

2. When and to what degree will depository library materials (especially census data) be distributed via CD-ROM or other machine-readable media?

3. What technological developments will affect patrons' remote electronic access to depository libraries?

4. What software and data structures will be required for electronically disseminated census data, to facilitate rapid and effective searches and retrieval?

5. What will be the sponsorship and impact of standardization efforts to facilitate network access to federal government information?

6. To what extent will anti-trust concerns inhibit development of data integration protocols and telecommunications software necessary for widespread network access to federal government information?

7. How will the distribution of government information be controlled, under whose auspices and with what objectives?

8. How will data integrity be maintained without impeding widespread electronic dissemination of information?

9. Which sponsors of information production, dissemination and use will support high technology access, under what conditions and with what goals?

10. What are the prospects that Congress will choose to privatize depository library distribution? What impact would that have upon the quality, quantity, availability and cost of Census Bureau information?

11. What will be the minimum skill levels required of users and depository librarians in accessing electronically disseminated information?

12. To what degree might user fees and other costs of accessing electronically disseminated information disenfranchise individuals?

13. What impact will changes in work force composition and employment arrangements have upon the types of census information sought by users?

Overview of Future Trends

The Economy

Although estimates vary widely, some projections forecast a period of modest economic prosperity for the United States in the next two decades, including a strong rise in "knowledge industries."

Demographers predict an enlargement of the middle class, with fewer very poor or very wealthy. Cultural homogenization is anticipated, despite an increase in non-English speakers and the swelling ranks of citizens over 65, with each group having its own unique perspective and needs (Cetron, 1988).

Information consumption may be influenced significantly by an increase in middle class affluence. The incomes of middle-aged citizens will rise about three-fourths. One-third of middle-aged households will have annual incomes of $50,000 or more, in constant dollars (New American, 1986). More sophisticated, better-educated consumers with work experience will have disposable income for travel, leisure and luxury. Spending will continue to shift toward service industries (Cetron, 1988).

Education and Training

Four percent of the labor force may be in job retraining programs in the coming decade (Cetron, 1988). There may be more rigorous educational standards at all levels and a greater concern for human rights and personal freedom (Caddy, 1987).

Work Force Composition

By the Year 2000, manufacturing will employ only nine percent of the labor force, with services taking 88 percent, partly because productivity in automated industries may increase fivefold. Seventy percent of U.S. homes may have computers in 2000, facilitating the potential for widespread remote access to federal government information (Cetron, 1988).

Changes in the work force may be a key influence during the next decade. The mandatory retirement age may be 70 by the Year 2000. Union members will comprise less than 10 percent of the labor force (versus 29 percent in 1975 and 18 percent in 1985). The work force will be dynamic, with people changing careers an average of every 10 years. A shortage of low-wage workers will force businesses to automate and to seek foreign workers. The ranks of the self-employed will grow at a faster rate than salaried workers and more mid-career professionals will become entrepreneurs (Cetron, 1988).

Some expect a crisis of consumer confidence in the U.S. technological infrastructure as widespread hacker activity plagues computer networks (1988 Ten-Year Forecast, 1988).

Also anticipated are a decrease in the divorce rate, an increase in marriages and family formation, and a heightened role for religion. Do-it-yourself activities will be popular, because a 32-hour work week will create more leisure time and due to the high cost of services. Protracted adolescence may be more common, although there will be far fewer young people than at present. A decrease in the size of federal government will be accompanied by growth in state and local governments (Cetron, 1988).

Knowledge Industries

The United States is becoming a postindustrial society. In such a system, telecommunications and computers are vital to the exchange of information and knowledge (Bell, 1978). Multimedia information networks permeate everyday life. New information and processing devices increase productivity, despite initial retraining losses. Knowledge industries grow in importance.

As the central role of information accelerates, major policy issues will include privacy, the part government plays in information dissemination, intellectual property rights and functional literacy. Neural networks (combinations of electronic and photonic circuits used in optical computers) may be one of the key new technologies. The science of integrating diverse telecommunications and computing systems may become an important driver of technological innovation (Bezold and Olson, 1986).

Information Technologies

Despite the increasing use of computers in a wide variety of applications, there also may be an increase in the need for paper. Tenner (1988) reported that from 1959 to 1986 U.S. consumption of writing and printing paper increased 320 percent while real GNP rose only 280 percent. He believed that electronic information supplemented rather than replaced paper, and noted that using paper is more efficient, legible, and secure than working with monitor displays. Tenner also predicted that increases in the number of office workers will cause a corresponding growth in the use of photocopiers and facsimile machines (which use paper).

Massive increases in storage technology will take place, with commercial system capacities in the hundreds of megabytes. Optical disks (some erasable) will emerge as the medium of choice in applications such as census data retrieval that require high storage capacity, fast access, removability, non-contact recording, and long life. Engineers predict that current disk capacity will increase tenfold and that the 600 megabyte disk that now sells for $200 is likely to cost only $25-50 (Freese, 1988).

Knowledge in Postindustrial Society

The Economics of Information

Social organization will be shaped by intellectual technology in postindustrial America. Since information and knowledge are not depleted in the sense that goods are in an industrial economy, knowledge will be considered a social product. Its cost, price and value will be assessed in a way vastly different from that for industrial goods, in accordance with what is known as the "Knowledge Theory of Value" (Bell, 1978).

Knowledge, even when it is sold, remains with the producer. It is a "collective good" — once it has been created, it is available to all. There is little incentive for any single person or enterprise to pay for the production of knowledge unless a proprietary advantage (such as a patent or copyright registration) can be obtained. Thus, government policy in regard to intellectual property and contractor marketing of publicly funded products will be key in the management of future information dissemination technology.

Bell (1978) believes that a reduction in incentives for individuals or companies to produce knowledge will cause the responsibility for and costs of satisfying information needs to fall to government. Whether information dissemination is "privatized" and in what manner may affect the availability of that information significantly.

The first infrastructure placed in service by industrializing economies is transportaion. Fully industrial systems concentrate on energy utilities. Societies entering the

postindustrial phase need advanced telecommunications. Therefore, the major technological problem for America in the next decade will be emplacement of an appropriate digital information network, carried over fiber optic cable (Bell, 1978).

Social Impacts

The U.S. as postindustrial state is likely to have a vastly different social structure than at present. Bell (1978) believes that this new order may be characterized by:

1. Centrality of theoretical knowledge as the basis of innovation.

2. Creation of new intellectual techniques to engineer solutions to economic (and even social) problems.

3. The spread of a (technical and professional) knowledge class.

4. The change from goods to human services.

5. A change in the character of work (people must learn to live with one another, since interaction among groups will be key).

6. The employment of women in expanded human services.

7. Science as the societal standard bearer.

8. Political units comprised of either vertical organizations of individuals into scientific, technological, administrative, and cultural centers, or of institutions arrayed as economic, government, university, or social complexes.

9. Meritocracy (an emphasis on education and skill).

10. Scarcities of information and of time.

11. The economics of information.

Bell (1978) also believes that information by its nature is collective, not private. In postindustrial America, the optimal social investment in knowledge may require that we follow a cooperative strategy to increase, spread and use knowledge.

A more pessimistic view of the impact of advanced telecommunications technology is taken by Eldredge (1978). Hebelieves that new technology will:

1. Will be highly beneficial to some segments of society, but detrimental to others.

2. Will have a positive impact primarily in the middle-class suburbs, with a negative impact in central cities.

3. Will not be properly understood and regulated until considerable damage has been done in major urban development.

4. Will reduce the economic viability of the central city by accelerating delocalization of business and commerce.

5. Will affect the service sector most, because its processes involve paper transactions that are particularly sensitive to technological substitution.

The Depository Library Program

It is within this context of an increasingly central role for the development and dissemination of information that we consider future management of the Depository Library Program.

The federal government has a long history of providing increasing amounts of information to the public as part of its responsibility to maintain an informed citizenry.

In particular, substantial volumes of information ranging from census data to contract studies of government activities to congressional hearings have been disseminated in a network of some 1400 libraries as part of the Depository Library Program. Under the program, government-printed material is distributed to a limited number of regional depository libraries. Additional "select" libraries also archive some subset of these materials for their patrons.[2]

Recently, census data have been distributed to a few depository libraries on CD-ROM in an effort to assess the medium's potential, as well as to gauge user reaction.[3] Bureau of the Census material also is available to institutions and to the public at State Data Centers operated specifically for that purpose.

Discussion of Key Research Questions in Terms of Findings

Elite interviewing and the literature review for this study identified five major areas that may affect future dissemination of census data as part of the Depository Library Program. These areas were selected based upon the emphasis they were given by interviewees and upon the author's experience in information systems design.

1. Technology

2. Regulation

3. Funding

4. Access

5. User Demand

Technology

Leading Edge Technologies

What will be the leading edge information technologies in the first decade of the next century?

The media used in dissemination of federal information in the Year 2000 may be a mixture of optimized current-day technologies (McGee, 1990).

Data input, the bane of current full-text efforts, may no longer be a problem. Research during the past few years has made it possible for many agencies (e.g., the Air Force) to use improved optical scanning devices to make large amounts of printed text machine-readable (McGee, 1990). Once standards are established for sharing scanned input, an enormous amount of digital data will be available, although coordination and retrieval problems will need to be addressed.

Information dissemination also may include specially engraved static memory chips for advanced personal computers. Response time typical of such devices would be significantly shorter than with mechanical access systems — an important consideration for complex, natural language queries of large databases. Inexpensive, high-capacity chips will be available. (In a recent demonstration, IBM technicians wrote patterns at single-atom resolution using a scanning tunnelling microscope.[4]

Thus far, vast sums of money have been spent on advanced technologies without any real understanding of how people might interrelate with these devices (Weiner & Brown, 1989). Current improvements center on the ability to gather, store, and catalog information.

Much information dissemination in the Year 2000 will be via fiber optic cable, which due to its enormous capacity will compete effectively with relatively limited capacity direct-broadcast satellite transmission in many real time access applications. Much of the internal transmission capacity of most telephone companies already has been converted to fiber, and a number of large businesses have access to fiber networks. All subscribers are certain to be fiber-connected in 25 years, with many equipped during the 1990s (Weinstein & Shumate, 1989).

There will be an explosion of communications options as the fiber optic infrastructure is installed, making possible single transmission, multiple-service options such as on-line catalog ordering and public opinion polling (due to the signal capacity of fiber optic cable). Today, the Integrated Services Digital Network (ISDN) makes possible voice, data and image transmission over the same phone lines. ISDN is likely to be deployed widely by the end of century, making possible high fidelity audio and five-second per page facsimile. Early in the next century, ISDN will give way to the Broadband Integrated Services Digital Network (BISDN) — an intelligent networks supporting high quality, simultaneous and on- demand video services, home telemetry, and high speed data and image communication among faxes, workstations and computers. The Telecommunications Network for the Deaf (TND) will convert speech into text and vice versa to assist handicapped subscribers. These and other protocols may be tested soon in the NREN high-speed network already in use at many libraries.

CD-ROM and on-line technology early in the next century will assist the user in sorting through the thicket of available information by means of "information grazing": an intelligent, user-selectable filtering system capable of passing only items of interest in order to manage information overload.

Machine-Readable Media

When and to what degree will depository library materials (especially census data) be distributed via CD-ROM or other machine-readable media?

It is already within the means of depository libraries to provide public access to census data via CD-ROM. Recently, the Bureau of the Census has been engaged in a test distribution of files on CD-ROM to depository libraries throughout the country (J. Stratford, personal communication, February 21, 1990).

However, the test discs have not been received well by all depositories. Data on one disc were formatted in a different pattern for each file, making it necessary for users to apply different techniques to access each dataset. These variations in format have decreased the usefulness of the disc (K. Chiang, personal communication, April 12, 1990). Also, a virus was distributed on the 1988 County and City Data Book CD-ROM, though prompt action by the Depository Library Program appears to have averted serious problems for member libraries (Harm, 1990).

Despite such initial problems, CD-ROM seems ideally suited to dissemination of census information which by

its nature contains large amounts of data. In the mid-1980s, Grolier placed its 21- volume, nine-million word Academic American Encyclopedia on a single compact disc. Eighty percent of the disc's capacity remained unused. In addition to enormous space savings, use of CD-ROM made possible high-speed keyword searches (Cornish, 1985).

In the future, CDs will not necessarily be distributed to all institutions in the Depository Library Program. Much in the way that only regional libraries receive full dissemination now, some expect discs bearing federal information to be distributed only to regional depository libraries. Select libraries could request discs on loan as required by their patrons. (McGee, 1990).

Remote Electronic Access

> What technological developments will affect patrons' remote electronic access to depository libraries?

CD-ROM will by no means be the only electronic access to federal information in the years ahead. Combinations of new and existing technologies will facilitate the widespread availability of census data early in the 21st Century.

Integration of facsimile devices with home entertainment centers will allow users to receive hardcopy on demand of a wide range of federal government information, if on-line access to depository databases is allowed. Installation of fiber-optic cable to homes and businesses will make possible a dynamic, interactive census process in which users not only access more current database information but also can register data about themselves more easily and more frequently than is possible with current, paper-based census techniques. The Year 2000 may be the last "traditional" census, as citizens enter the data collection, processing, analysis and dissemination loop more actively.

Such dynamic census systems would raise issues of quality control, user cooperation, and prevention of abuse by commercial marketing interests. Extension of on-line access to the full range of depository data also may call into question the need for regionally distributed depository archives.

Most depository libraries already are involved in networks. There are 20 regional networks, plus some CD-ROM distribution. Some federal information specialists feel that depository library information should be on-line, if only to reduce the expense of storing information that is used infrequently. Now, it is enormously expensive and wasteful to print and distribute materials that few if any patrons use: "Much of government information is a record going nowhere." (McGee, 1990).

Regardless of storage capacity, it still will be necessary to keep truly unused information from clogging the system. It has been suggested that librarians help define the characteristics of a filter to be used in deciding what depository information should be part of on-line and CD-ROM databases. Information use could be evaluated at regional libraries, or such institutions could delegate responsibility to subject specialists. However, many librarians resist the filtering of information, preferring to make decisions on a case-by-case basis rather than allow data to be withheld at centralized distribution points (McGee, 1990).

Software and Data Structures

> What software and data structures will be required for electronically disseminated census data, to facilitate rapid and effective searches and retrieval?

The complexity and volume of federal information in all formats will require sophisticated indexing and retrieval software, but existing government databases are seriously lacking in such tools (McGee, 1990). Software used currently to access the test CD-ROMs distributed by the Bureau of the Census also may be inadequate ("Until there is adequate access software, the CD is not much of an improvement on a stack of microfiche." — K. Chiang, personal communication, April 12, 1990).

Efficient software will need to be developed to manage the volume of machine-readable information available in the future. When the mass of retrievable information is more than the brain can process effectively, the result is not faster decision making, but instead a delay in or even abdication of decision making (Weiner & Brown, 1989). Clarke (1985) noted that without adequate indexing, many users would be completely overwhelmed by a virtually limitless selection of information resources, and chose to select nothing. He also pointed out that

> *With the latest techniques, it would be possible to put the whole of human knowledge into a shoe box. The problem, of course, is to get it out again; anything misfiled* would be irretrievably lost.

Within the next decade, artificial intelligence will be applied to such problems (Diebold, 1985).

Some data in future federal databases may need to be in a format suitable for manipulation by spreadsheet and statistical programs such as Lotus 1-2-3 or SAS (McGee, 1990). The widespread availability of relatively inexpensive bitmapped displays, faster processors, and a demand for three-dimensional graphics and photographic-quality images may force some network databases to be stored in

tokenized or vector formats to be regenerated at user workstations. (Graphics regeneration has been available for some time on more expensive minicomputers such as the DEC MicroVax II, using the ANSI/ISO-standard Graphical Kernel System.) Such formats allow specialized processors in users' machines to plot images from line endpoint data rather simply displaying full-screen files downloaded from host computers. This capability would accelerate display speeds significantly, partly due to new data compression techniques and the capacity of fully digital, high speed networks to download enormous amounts of data.

Regulation

The history of innovation is replete with examples of new technology that was not implemented to its potential because of social or political factors that frustrated its use. Future data dissemination technology could suffer a similar fate if issues of standardization, anti-trust, control of distribution and data integrity are not addressed adequately.

Standardization

> What will be the sponsorship and impact of standardization efforts to facilitate network access to federal government information?

The architecture of future data access systems and the standards coordination required for integration of diverse computing hardware are key concerns that will need to be addressed as technology makes large volumes of federal data available in machine-readable format. As Clarke (1985) observed,

> *Another problem is to decide whether we mass produce the shoe boxes [databases], so that every family has one, or whether we have a central shoe box linked to the home with wideband communications.*

Information specialists at the Library of Congress believe that a single, coordinated, national database with remote access is not likely. Interface standards will make possible an extensive, distributed architecture populated with vastly different hardware in a highly interconnected system. (Substantial interconnection already exists among NTIS, DOE and Medline.) They feel that the centralized database concept is obsolete. In this view, future on- line access to federal information would more likely be via highly distributed selective centers, linked to each other using standard protocols (Bortnick & Relyea, 1990).

Establishment of standards for such on-line networks is a key concern, since integration of computers from a multitude of vendors operating under vastly different operating systems may be involved. Equipment already in place at user sites could be linked to provide reasonably economical service, if sufficient standards were developed by government and applied as part of the access system. Standards may not even be imposed by the federal government, but instead by international entities, owing to the substantial interconnection even now between federal government data stores and overseas sources. European networks have tended to be further advanced than those in the U.S., therefore are more likely to drive any movement toward standardization (Bortnick & Relyea, 1990).

If an institution could not link to the system using existing hardware because its plant was antiquated, funding could be requested from corporations, the Department of Commerce and/or the National Science Foundation to bring the site to minimum levels for participation in the system. Otherwise, it would be presumed that institutions could tap into the on-line system with existing equipment and software, if they met standards (Bortnick & Relyea, 1990).

At present, standardization even within the federal government is difficult. Many agencies have material in electronic form, but there is no coordination of data formats or machine compatibility. Most electronically stored data are not available outside their host agency. Although the Office of Management and Budget (OMB) or General Services Administration (GSA) may have the authority to force coordination of data formats throughout the federal government, it remains to be seen whether that authority will be applied, and with what result (Powell, 1990).

Anti-Trust Concerns

> To what extent will anti-trust concerns inhibit development of data integration protocols and telecommunications software necessary for widespread network access to federal government information?

Traditionally, federal anti-trust law has prevented firms in competition with each other from cooperating in ways that may be necessary for development of standardized protocols, equipment and software for the fully integrated information systems of the future. However, anti-trust law has changed in recent years to allow cooperative research and development among firms, largely in response to competitive market pressures from overseas (where such cooperation is common) (Bortnick & Relyea, 1990).

There are proposals in Congress to extend this liberalization to include product development and even production. The availability of seamless networks for user access to on-line federal information would be affected

significantly by greater cooperation among service suppliers, if they were not constrained by anti-trust regulation.

Federal anti-trust enforcement has been dormant in recent years, in part because there have been few substantive changes in business activity from old, established patterns. However, the new information technology constitutes just such a substantive change, implying the need for changes that would facilitate cooperative development of technologies — especially protocols and access software — that would facilitate interconnection (Bortnick & Relyea, 1990)

Control of Distribution

> How will the distribution of government information be controlled, under whose auspices and with what objectives?

Telecommunications policy centers around who provides services, under what conditions, and at what prices.

A jurisdictional struggle is under way now among the Government Printing Office (GPO), OMB, National Technical Information Service (NTIS) and others over control of federal information dissemination. The GPO is funded by the legislative branch; clashes take place between the executive branch and Congress over information policy. Each side has its allies in Congress. (Both the House and Senate have committees with oversight authority for each agency involved in the policy debate.) Paperwork reduction is the concern of the Senate Committee on Government Affairs[5], statistical policy is set by the Bureau of the Census, and the Depository Library Program is managed by the Joint Committee on Printing. Commerce and Science committees also are involved, because of technology issues. It is difficult to make policy with such fragmentation (Powell, 1990), and just as difficult to implement it because of the tug of war among the actors in their attempts to influence appropriations.

House bill 3849 (introduced in January) is one manisfestation of the struggle. This legislation attempts to prevent public monies from being used by the executive branch to generate and distribute information products and services without involvement of GPO. It broadens the legal definition of "documents" to include information products "in any tangible format, medium or substate", and attempts to interpose the Superintendent of Documents between the depository libraries and any government agency that might issue information products (Government, 1990).

This oversight problem must be resolved before the depository system can upgrade, regardless of the benefits of technology. Some observers feel that the Depository Library Program will need to be removed from GPO auspices before its technological potential can be fully realized, in part because the GPO work force is hesitant to diversify from traditional media (Bortnick & Relyea, 1990).

Some parts of the library community feel threatened by the new attempts to centralize and standardize the new technology. In 1985, OMB's Office of Information and Regulatory Affairs issued Bulletin A-130, which dealt with dissemination of federal information in electronic format and was part of the effort to reduce the volume of government printing (Powell, 1990). Some librarians felt the policy would undercut the depository library system (Powell, 1990). Initially, the cost of advanced technology may buttress such resistance by those who favor hardcopy dissemination (McGee, 1990).

Questions of access to federal information networks also need to be addressed before practical implementation of rapidly developing technology. Some within the federal community feel that capabilities such as full-text retrieval should be available to all citizens, not merely to depository libraries (McGee, 1990).

Some concern has been expressed within government regarding regulation specifically of the on-line dissemination of federal information due to provisions of the Export Administration Act (PL96-72). They feel that the competitiveness of American industry might be impaired if information were available to foreign competition. Also, DoD is concerned about the "mosaic theory" potential of widespread, machine-readable federal information (i.e., what new knowledge can be gained from machine correlation of public information).

Therefore, policy will need to be made regarding whether on- line information should be available only to U.S. citizens. Restriction may not be feasible regardless of policy because of the difficulties of controlling information transfer in an environment of total interchangeability among data formats. Release in one format would be tantamount to release in all others (Bortnick & Relyea, 1990).

Further frustrating attempts to limit overseas access to federal information is the fact that a great deal of information in existing federal databases (e.g., NTIS or DOE) is from foreign sources. It is unlikely that overseas concerns would be willing to continue providing substantial amounts of information to a U.S. database if they were not allowed in turn to access that database.

In any case, controls cost money which might not be forthcoming, depending upon the perceived importance

of such restrictions compared to the perceived need for convenient and widespread public access to federal government information.

Data Integrity

> How will data integrity be maintained without impeding widespread electronic dissemination of information?

Data in a freely accessed, on-line system could be copied, modified then redistributed, without the knowledge of the issuing agency. Original data may even be subject to accidental or malicious corruption. It might be difficult to protect important data whose accuracy would be presumed to be the responsibility of the Bureau of the Census (Bortnick & Relyea, 1990).

With the advent of new technology, however, a different model of responsibility might be in place. In the past, a publisher may have been expected in certain circumstances to notify the trade media (e.g., Publishers Weekly) about serious errors discovered after a book was distributed. Easily modified works such as those marketed in three-ring notebooks or with spiral bindings could have been distorted in the field, totally unknown to the original publisher. In the analogous future situation with machine-readable material, the aspect of the technology that makes data easy to corrupt also makes unauthorized changes relatively easy to correct, if intrusions can be detected.

Funding

Sponsors

> Which sponsors of information production, dissemination and use will support high technology access, under what conditions and with what goals?

Funding is a natural barrier to the widespread dissemination of information. Who should be responsible for capital equipment and telecommunications costs incurred if patrons or depository librarians are to have on-line access to census data?

In all probability, future funding for information dissemination would follow existing models. These include (1) a prototype database access system sponsored by the Library of Congress, (2) the regional supercomputer program, and (3) the Federal Information Center Program.

In the Library of Congress model, 12 libraries, corporations and agencies across the U.S. use their own equipment and staff, and pay their own telecommunications charges. The Library maintains the database and provides access to its computers (Bortnick & Relyea, 1990). The Internet communications network operates on a similar model. Universities across the world provide their own equipment and software; the U.S. government maintains the relatively small interconnection "backbone" of the system.

In the regional supercomputer program model, access to information is restricted to institutions that provide financial and support for the system. Contracts between government and the universities hosting the centers include specific provisions for use and funding. Centers typically solicit financial support from two or three distinct sources (Bortnick & Relyea, 1990). Under this model, depository libraries might maintain regional CD-ROM or on- line database centers which loan discs or provide logon accounts to institutions that pay to support the system. Whether the subscribing institutions in turn charge their patrons for access is a major policy option.

Some at the Library of Congress feel that the Depository Library program should be subsumed by the GSA's Federal Information Center (FIC) Program, making it the central point from which to access many types of integrated media. These Centers are not funded solely by the legislative branch, but instead are supported by many sources under a "plurality of funding" concept that is more acceptable politically (Bortnick & Relyea, 1990). Under the FIC model, a private firm (Biospherics, Incorporated of Beltsville, Maryland) manages regional offices which field telephoned inquiries regarding a wide range of government services, programs and regulations. By terms of the contract, Biospherics is is required to provide access to the hearing and speech impaired (Federal, 1990). Under this model, depository library materials presumably would be accessed on-line using regional FIC database systems.

Depository libraries are not very high on the political agenda in Congress, since their funding comes directly from the legislative branch budget. Many legislators believe that government should provide the information, but not the means for its distribution. In the future, the federal subsidy for information dissemination might be limited to the current cost of paper distribution (Bortnick & Relyea, 1990).

Telecommunication charges for the dissemination of federal information may be regulated in the future by policies such as the FCC Open Network Architecture (ONA) initiative. ONA's intent is to prevent monopoly control by local telephone carriers of connections to national networks. The initiative requires modular pricing of discriminible services, to promote competition among suppliers.

If on-line access to depository library information is to be provided in the home, then charges for terminal electronics and for the in-ground installation of fiber-optic cable must be comparable to the total cost of a typical telephone line installation today: $1,000 to $1,500 per home. Network operation also must be economically feasible. That it may be is demonstrated by research at Bellcore which has resulted in experimental models of broadband, digital networks (Weinstein & Shumate, 1989).

During the more than a century that the depository library program has provided public information to participating institutions, the government has funded the printing and dissemination of materials. Changes in perception of government's proper role — and in its ability to support initially expensive dissemination technology — make it doubtful that future funding will come from Washington alone.

Funding expanded network access in particular is a serious concern. Some federal information specialists believe that future on-line systems will need to be fee-based. Federal libraries already pay for on-line access, in a manner similar to that used to charge commercial customers for access to commercial databases such as Dialog.

Despite these concerns, some analysts believe that basic information utilities of the future will be economical to the point of being taken for granted. Their concern is that such a cheapening of information will risk potential devaluation of product quality and intellectual creativity (Diebold, 1985). It should be noted, however, that low cost does not necessarily translate into easy access. Regardless of cost, there still will be a need to filter the vast amounts of information that will be available in order to retrieve only what is needed.

Privatization

> What are the prospects that Congress will choose to privatize depository library distribution? What impact would that have upon the quality, quantity, availability and cost of Census Bureau information?

There are some who feel that OMB's Circular A-130 (and the corresponding requirement for A-76 studies)[6] is too favorable to the private sector, because it urges agencies to contract for information services (Powell, 1990). This sentiment underlies HR 3849, the Government Printing Office Improvement Act of 1990, which if passed would require that depository libraries apply through the Superintendent of Documents for any government information product, and that such application include the specific cost sharing arrangements proposed among users, the depository libraries, the issuing agency and federal appropriations (Government, 1990).

Joint government-industry initiatives may be inevitable, not only because of federal funding constraints, but also because private industry holds advanced search and indexing software necessary for the management of comprehensive, interrelated data stores. Possession of such tools by firms such as Dialog and Nexus place them in an excellent position to bid contracts for on-line or CD-ROM access to federal information (Bortnick & Relyea, 1990). Joint initiatives are prompted in part by the concern of private industry regarding "unfair" competition with the government if federal information is placed on-line at subsidized rates.

The declining federal budget also may prompt partnerships with private industry in the future. Federal outlays will be a declining proportion of GNP for the rest of the century, with the growth rate of the budget steadily declining (New American, 1986). In recent years, nearly 30 percent of federal spending has gone to pay old-age benefits to the 11 percent of the population currently over 65. Far more will be receiving benefits in the future, as the overall U.S. population ages. In 1986, interest on the national debt (the fastest growing portion of the budget) accounted for 18 percent of all federal spending (Longman, 1988). These and other pressures on the federal treasury may force future information dissemination to be self-sufficient, by means of the involvement of commercial partners.

Access

Skill Requirements

> What will be the minimum skill levels required of users and depository librarians in accessing electronically disseminated information?

Projections of median education and skill levels in the future point to a wide divergence in patrons' basic understanding of new technology and the information that it will provide. A dichotomy is developing between a small, undereducated and underskilled younger generation and a larger, educated and job- experienced retirement cohort (Longman, 1988).

Longman also reported that today's younger generation is not only comparatively small (due to low birth rates in the last two decades), but also that an alarming proportion of youth lack the basic skills that employers require. Only 30 percent of today's 17- year-olds are classified as "adept" readers (i.e., competent enough to go on to college or to cope with business environments). In the early 1980s, a 12-nation study found that U.S. average comprehensive scores on seven school subjects always were in the lower third.

Since 1973, the poverty rate among Americans under 18 has increased more than 50 percent. Each year, hundreds of thousands of young people reach working age without the basic knowledge they need to learn even the simple skills necessary for success in an entry-level job. The implications for their use of depository libraries and access to census data are serious: either data access must be simplified to the extent that the unskilled can retrieve needed information or depository librarians will experience a significantly increased workload acting as intermediaries for such patrons. Otherwise, the unskilled will become disenfranchised (see below).

Nearly all economists agree that the industrialized nations are moving toward information-driven, rather than energy-driven, economies in the next century. The intellectual skills of the labor force in such systems will become increasingly important to maintaining a comparative advantage in international trade. Skills deficiencies among younger patrons may create a frustrated, information underprivileged class with comparatively little real political power.

Fortunately, placement of CD-ROM devices in schools and the development of improved interfaces may make possible gradations of technological user-friendliness in the next 10-20 years, if either funding or equipment and software donations to schools can be arranged. Literacy either in the traditional sense or in a technological sense will decrease in importance. For example, the Library of Congress reading room reconstruction currently under way will include installation of touch screen terminals. It is hoped that such devices will help non-technically sophisticated patrons access the collections (Bortnick & Ralyea, 1990).

There is some question, however, whether library staff will be skilled enough with the new technology. Part of the startup costs associated with CD-ROM or even on-line access as an integral part of the Depository Library Program may be the training of staff so they can assist others in using the systems.

Disenfranchisement

> To what degree might user fees and other costs of accessing electronically disseminated information disenfranchise individuals?

Since the founding of the depository library system, it has been presumed that participatory democracy dictated widespread dissemination of public information. However, the increasing cost of the distribution effort is causing this premise to be reexamined. Some information specialists question the societal costs and benefits of universal information suffrage (Bortnick & Relyea, 1990). They ask

> 1. Is widespread dissemination of public information essential to the preservation of an acceptable level of shared values among citizens?
>
> 2. Would acccess charges imperil that acceptable level of shared values and create an information underclass?

Future technology may provide the flexibility to manage funding problems, making possible the continued widespread availability of information.

Expensive, on-line access is not essential. More economical CD-ROM media will be cheaper, and enhanced video systems costing no more than a TV does today will provide cheap, multimedia access (Bortnick & Relyea, 1990). To the extent that direct broadcast satellite reception and public access cable channels are available, those technologies may serve to democratize information dissemination. (Direct broadcast is in use over India today. As minimum antenna diameter shrinks to less than a meter, home satellite reception may become much more widespread than it is today. Also, within the decade there may be enough cable capacity that almost any group that wants its own channel can have it for special broadcasts (New American, 1986).)

However, if market forces are allowed free reign, history suggests that large businesses and institutions will have disproportionate access to the new technology (and to the information it provides), because they are most able to afford capital and staffing costs and will have the most influence in information systems development. If funding continues to be a problem, younger individuals may be disadvantaged in comparison to older users, companies and universities.

User Demand

> What impact will changes in work force composition and employment arrangements have upon the types of census information sought by users?

The Aging Population

The number of Americans over age 75 will grow almost 35 percent by the year 2000. The U.S. population 65 or over in 2035 will be 22 percent, vs. 12 percent in 1985 (Haub & von Cube, 1987). Currently, one of three Americans is between 27 and 42 (the Baby Boom generation). The oldest boomers are just 19 years away from reaching the current average age of retirement (Longman, 1988). One consequence of changes in age distribution will be an increase in two-generation geriatric families during the 1990s — adult children in their

60s and 70s caring for parents in their 90s (Outlook, 1989).

The population of retirees will be characterized by a higher average level of education, a friendliness toward business and free enterprise, entrepreneurism, a skepticism toward big government, an inclination to regionalism, and a preference for decentralized regulation (New American, 1986).

Between 1970 and 1980, life expectancy at 65 increased more than nine percent (life expectancy at birth increased only three percent, thus the fastest growing segment of the population is the age group over 80). Both the absolute number and the proportion of the population under 25 are declining. By 1995, people 16 to 24 will be only 16 percent of the population (they were 25 percent in 1980) (Longman, 1988). Serious attempts to slow the human aging processes and prolong life expectancy may begin within the next decade (Outlook, 1989). By the Year 2000, life expectancy will be 72.9 years for males and 80.5 years for females (New American, 1986).

Also, the physical design and environment of America (information services included) will start to change in the 1990s to accommodate a middle-aged and older population. (For example, traffic lights will change more slowly, allowing more time for less-agile people to get across intersections).

The graying of America may have a significant impact upon the types of information sought by individuals (and institutions operating on their behalf). Seniors are likely to have a special interest in census data because of its link to medical and social security benefits, its bearing upon investment and savings decisions, and its potential as ammununition in lobbying Congress. They will have more time to focus upon and use available data, and will constitute a trained, educated clientele able to make drastically increased demands upon the system.

Changes In The Work Force

By the year 2000, 95 percent of all jobs will be in service industries that require workers who are familiar with computers and other information processing technologies.

The U.S. may move toward a dual economy, where professionals, scientists, engineers, technicians and other skilled employees are on one end of the spectrum and a large number of blue-collar workers, clerks, and service workers are on the other (Lamm, 1985).

Telecommuting and other flexible-place, flexible-time work schedules will become increasingly common as employers acknowledge modern realities such as single parent households and the stress of urban commuting (Outlook, 1989). Changes in information need, therefore in the use of depository library material, are certain to follow such changes in lifestyle. Hobbies, travel, and intellectual pursuits may become higher priority for some, while others seek inforamtion in self-help, quality of life areas.

Retirement may become a thing of the past, as seniors remain in jobs at all levels in the workplace. Many people will return to work after a sabbatical, act as consultants or become "senior apprentices" to learn new skills for a second or even third career (Outlook, 1989).

Multilingual Services

Work force and patron demographics also will drive the development of multilingual information services in the next century's depository program. The population growth of 12 percent anticipated for the next 15 years will result almost entirely from high level immigration and from the higher-than-replacement birthrate of new immigrants (mostly Hispanic). In the mid-1990s, Mexico's proportion of young job-seekers is due to about double, while the kinds of entry-level jobs they seek at home will shrink drastically. The result will be enormous pressures at the border which will not be entirely unwelcome, as U.S. employers struggle to deal with a shortage of American-born youngsters in the labor force. The user community of the future will include a steadily rising percentage of racial minorities, many using English as a second language. By the Year 2000, 11 percent of the population will be Hispanic and four percent Asian; by 2020, Hispanics will be 15 percent (New American, 1986).

Patrons whose first language is not English are likely to need information in their native language, regarding culturally specific subjects at variance with those sought currently by typical depository library patrons. Immigration, employment and family data may be more important to this group than to the population at large.

The growth in non-English patrons also may add impetus to the development of non-culturally specific interfaces. Even with new storage technology, simultaneous storage of text in Asian and Hispanic languages as well as English could be prohibitively expensive, as demonstrated by the Canadian experience with French and English. However, automatic translation systems being developed as part of the next generation of computers may help solve the multilingual problem. By the year 2000, computers with automatic language translation and voice-synthesis capabilities may enable people to speak in one language that listeners will hear translated into another language (Outlook, 1989).

Key Issues Requiring Public Debate

A number of key issues settled out of the literature

review and interviews conducted for this study. These are issues that either policymakers, the Bureau of the Census or the depository libraries need to address before developing technology forces less than optimal solutions.

Policymakers

> Should future information dissemination be oriented toward individual users or toward businesses and institutions?

Political and fiscal reality dictates that dissemination decisions weigh business and institutional concerns over those of individuals. These may include predominant formats, time of availability for on-line services, content of information disseminated, and fee structures. However, an effort should be made to ensure that institutions served by depository dissemination are reasonably responsive to individuals' requests, even if fees are charged for services rendered.

> Should joint ventures with private industry be pursued as a means of funding future dissemination in the face of a shrinking federal budget?

Fiscal reality may force government into partnerships with public database agencies, and to take advantage of advanced indexing and retrieval software and to defray data generation costs. However, in order to protect the public's right to access government information, some regulation of user fees charged by industry partners may be necessary.

> What policies should be adopted regarding intellectual property rights in data analysis, access software development and copyright protection?

Partnerships with industry will require special copyright provisions to protect the rights of government's partners, despite the fact that the products withheld would be generated in part with public funds. Contract provisions should state explicitly that key material is not "work for hire", allowing contractors rather than the government to retain copyright. Patents should be filed jointly by contractors and the government. These protections will need to extend specifically to indexing and retrieval software and network protocols developed in support of national on-line access systems. Government will no longer be able to claim its right to contractor source code and algorithms as a condition of working with private industry.

> How and where should advanced indexing and retrieval software be procured for access to machine-readable data?

If government enters into a partnership with private industry for database management, the vital access software will come from existing high quality products held by commercial firms. Otherwise, a major contracting effort will need to be staged to have suitable software developed for use with electronically disseminated federal information. Since technology will be evolving rapidly, the decision to develop software under government auspices would be a long term commitment of funds and manpower.

> What role should be played in coordination of federal information dissemination policy to eliminate fragmentation of jurisdiction over media, content and formats?

It is clear that a single, centralized agency is needed to coordinate electronic dissemination of federal information. A quasi-government entity including representatives from major actors such as the GPO, Congressional agencies and depository libraries could be organized under an Institute for Information Policy & Research (as described in HR744, 1985). The tendency to date has been for each agency to select its own standards and media without regard for activities elsewhere in government. Further, conflicting funding and advocacy situations within congress and the executive branch have frustrated standards development. The Bureau of the Census should mount a concerted effort to help resolve the jurisdictional confusion that currently exists.

Bureau of the Census

> How should demands for multilingual presentation beaddressed?

There are two politically acceptable options. Either data should be disseminated through the Depository Library Program in English and Spanish, or it should be formatted in a markup language compatible with automatic translation from English to a handful of user native languages. It may be necessary to disseminate in two languages until the technology is commonly available to perform the automatic translation.

> To what extent is the Census Bureau liable for ensuring the integrity of data disseminated in machine-readable formats?

The Census Bureau's liability for the integrity of electronically disseminated data is a question for serious legal consideration in a relatively new area of the law. A comprehensive study of legal issues related to data integrity and other aspects of the new technology needs to be conducted. These include tort liability for damages due to the use of inaccurate or corrupted data and product

liability in case dissemination includes destructive computer viruses or defective media which physically damage depository library equipment.

> Would the Census Bureau be accountable for invasion of privacy or threats to defense or industry confidentiality that might result from the ability to manipulate data in machine-readable format (the "mosaic" issue)?

This issue may either be a moot point, overtaken by extensive network integration worldwide, or a major impediment to widespread dissemination. The Census Bureau needs to examine not only its legal liability regarding privacy issues, but also its political defenses against corporate and DoD initiatives that are certain to be fielded <end indent both>as the technology makes machine manipulation of public data possible.

> To what extent should the Census Bureau be involved in establishment of network protocol and human interface standards both within government and within industry?

The history of technological innovation has proven that those who are not actively involved in standards setting activities incur both real financial costs and intangible political losses when standards drafted by others are implemented. Despite considerable expenditures of manpower and other resources, the Bureau of the Census must attempt to guide the establishment of protocol and interface standards.

> How should responsibility and costs be divided for creation and maintenance of on-line access networks?

When on-line access becomes feasible, the Bureau of the Census should implement the supercomputer centers model to create and support the system. Businesses and institutions should provide hardware and staff support, while government maintains network interconnects and polices protocol standards.

> Should the Census Bureau abandon the depository library program in favor of alternative means of data dissemination, or be a driving force in effecting a restructuring of the program in keeping with new information needs and dissemination technology?

The existing depository library program may need substantial revision (e.g., removal of select libraries, assessment of user fees, and collection specialization in terms of media and content). If required, this might be accomplished via a cooperative arrangement with federal agencies and other depository libraries. However, given that near future electronic dissemination is likely to be via CD-ROM rather than on-line database access, the distribution channels and procedures of the current program may be useful into the next century. If direct access by patrons using integrated networks or through data centers becomes predominant, Census Bureau participation in the depository library program may be abandoned.

Depository Libraries

> What types of training should be provided for depository library staff to better enable them to deal with the challenges of new technology?

Member libraries need to train staff in two key aspects of new technology application: (1) how to operate and maintain local hardware and software themselves and (2) how to best instruct and guide patrons in use of the technology. Neither task will be easy, since there may be significant differences among librarians in their degree of technological sophistication (and motivation). Preparation of a training cadre for the Depository Library Program should be undertaken immediately, funded jointly by the Bureau of the Census, Congress and the GPO. These master instructors in turn should brief library training officers who could provide ongoing familiarization at each Program site.

> To what extent should collection acquisition, operating and other funds be diverted to the purchase of hardware and software to support the use of electronically disseminated information?

It may be difficult to convince librarians to spend limited funds on equipment to support user access rather than on collection acquisition. However, in the long term funds spent on electronic equipment will result in more extensive and productive access to existing collections. Because of the density of electronically disseminated data, money invested in CD-ROM and associated printers and services will quickly expand a library's actual collection even as funds devoted to traditional acquisition decline. Expenditures for hardware, software and services related to electronic media should be a significant line item in each facility's budget.

Further Research

> Should the Census Bureau decide upon the medium and content of information disseminated based upon extent and type of use research?

Given the resistance of the library community and the seeming lack of space concerns for projected media, it would seem unwise to attempt limiting the kinds of

information disseminated through the Depository Library Program. However, it would be prudent to undertake a systematic and ongoing study of usage patterns in case budget or technological constraints make filtering necessary in the future. Patron preferences for specific media in accessing certain types of information should be evaluated.

Conclusion

This study was fielded under the presumption that government will be required to continue providing public access to federal information as part of its commitment to maintaining the informed citizenry that is central to participatory democracy.

The nature of that access, however, is entwined in a host of social, economic and technology issues that must be addressed promptly if the pace of change is not to overwhelm policymakers as well as information intermediaries and users.

Information will be central to the knowledge economy of postindustrial America. However, the population will be split between a relatively affluent, educated retirement community and a smaller unskilled, undereducated younger group less able to deal with sophisticated information access.

The generation and dissemination of knowledge will be dependent upon the degree of protection for intellectual property in an environment that features easy unauthorized copying of proprietary materials. The technology predominant at depository libraries will be CD-ROM, with on-line database services a distant second, used mainly for current updates of timely information.

Before emerging technology can approach its potential, problems of efficient indexing and retrieval software, hardware compatibility and protocol standards must be resolved. Perhaps key in this effort will be a relaxation of anti-trust regulation to facilitate cooperative research, development and manufacture by major players in the telecommuncations and computing industries.

Ultimately, the twin issues of funding and regulation underlie all concerns regarding future information technology. Given the declining resources of federal government, privatization and user fees seem inevitable. Privatization brings with it the prospect of reduced access to public information, and user fees the near certainty of disenfranchisement for a new underclass: the information disadvantaged.

These problems are not intractable. However, significant changes in the way we fund, generate, control and disseminate public information will be forced by technological change. If indeed those who do not learn from the past are condemned to repeat it, those who do not anticipate the future may be destined to live it amidst laments of what might have been.

References

Bell, D. (1978). The postindustrial economy. In J. Fowles, Handbook of Futures Research. Westport, Conn.: Greenwood Press, 507-514.

Bezold, C. & Olson, R. (1986). The Information Millenium: Alternative Futures. Washington: Information Industry Assn.

Bortnick, J. & Relyea, H. (1990, March 30). Interview at the Madison Building, Library of Congress, Washington, D.C.

Caddy, D. (1987). Exploring America's Future. College Station, Tex.: Texas A&M Univ. Press.

Cetron, M. (1988). Into the 21st century. The Futurist, 22:4, 29-40.

Clarke, A. (1978). Communications in the future. In J. Fowles, Handbook of Futures Research. Westport, Conn.: Greenwood Press, 637-652.

Cornish, E. (1985). The library of the future. The Futurist, 19:6, 2, 39.

Diebold, J. (1985). New challenges for the information age. The Futurist, 19:3, 68.

Eldredge, H. (1978). Urban Futures. In J. Fowles, Handbook of Futures Research. Westport, Conn.: Greenwood Press, 617-636.

Federal information center program (1990). [A background paper.] (Available from the GSA Information Resources Management Service, Washington, DC 20405).

Freese, R. (1988). Optical disks become erasable. IEEE Spectrum, 26:2, 41-45.

Government printing office improvement act of 1990. [HR 3849]. (Available from the Superintendent of Documents, Washington, DC).

Harm is averted by quick response to computer virus (1990). Administrative Notes, 11 (April 13), 1. Newsletter of the Federal Depository Library Program.

Haub, C. & von Cube, A. (1987). The United States

Population Data Sheet (6th ed.). Washington, D.C.: Population Reference Bureau. In Marien, M. (Ed.), Future Survey Annual (Item Nr. 8369). Washington, D.C.: World Future Society.

Hernon, P. & McClure, C. (1987). Federal Information Policies in the 1980's: Conflicts and Issues. Norwood, N.J.: Ablex.

Lamm, R. (1985). Megatraumas: America At the Year 2000. Boston: Houghton Mifflin.

Longman, P. (1988). The challenge of an aging society. The Futurist, 23:5, 33-37.

Marshall, C. & Rossman, G. (1989). Designing Qualitative Research.Newbury Park, Calif.: Sage.

McGee, Milton (1990, March 30). Interview at the Adams Building, Library of Congress, Washington, D.C.

The New American Boom. (1986). Kiplinger Washington Letter. Washington, D.C.: The Kiplinger Washington Editors, Inc. 1988 Ten-Year Forecast. (1988). Menlo Park, Calif.: Institute For the Future. Outlook '90 and beyond. (1989). The Futurist, 23:6, 53-60.

Powell, Elizabeth (1990, February 16). Interview at the Hart Senate Office Building, Washington, D.C.

Tenner, E. The Revenge of Paper. The New York Times, March 5, 1988, 27.

Weiner, E. & Brown, A. (1989) Human factors: The gap between humans and machines. The Futurist, 23:3, 9-11.

Weinstein, S. & Shumate, P. (1989). Beyond the telephone: new ways to communicate. The Futurist, 23:6, 8-12.

Additional Sources

The following sources were identified in the literature search for this study, but are not referenced in the final report.

Cornish, E. (Ed.) (1982). Communications Tomorrow. Bethesda, Md.: World Future Society.

Didsbury, H. (Ed.) (1982). Communications and the Future. Bethesda, Md.: World Future Society.

Dowlin, K. (1984). The Electronic Library. New York: Neal-Schuman.

Ferrarotti, F. (1986). Five Scenarios for the Year 2000. New York: Greenwood Press.

Gorman, M. (Ed.) (1984). Crossroads. (Proceedings of the First National Conference of the Library and Information Technology Assn., Sept. 17-21, 1983, Baltimore, Md.). Chicago: American Library Assn.

Naisbitt, J. (1982). Megatrends. New York: Warner Books.

Pasqualini, B. (Ed.) (1987). Dollars and Sense: Implications of the New Online Technology for Managing the Library. Chicago: American Library Assn.

[1] Presented at the IASSIST 90 Conference held in Poughkeepsie, N.Y. May 30 - June 2, 1990.
The author may be contacted at 4-290 Center for Science and Technology, Syracuse University, Syracuse, NY 13244-4100, or at D01DAYXX@SUVM.ACS.SYR.EDU.

[2] Hernon, P., McClure, C., & G. Purcell (1985). GPO's Depository Library Program. Norwood, N.J.: Ablex.

[3]Some information intermediaries believe it is important that federal information be made available via the latest technologies (J. Stratford, personal communication, February 21, 1990). However, problems with inconsistent file formatting and a lack of satisfactory retrieval software have made tests of experimental census distribution on CD-ROM less than a complete success (K. Chiang, personal communication, April 12, 1990). Problems of information policy making within a web of overlapping agency jurisdictions also have frustrated modernization efforts.

[4] Hudson, R. (1990, April 5). IBM researchers "write" with atoms on a metal surface. The Wall Street Journal, p. B4.

[4] Reauthorization of the Paperwork Reduction Act (1989). Hearings before the Subcommittee on Government Information and Regulation of the Committee on Governmental Affairs United States Senate (Senate hearing 101-166; Document 19-630). Washington, D.C.: U.S. Government Printing Office.

[6] OMB Circular A-76 mandated that agencies determine whether it would be more beneficial to continue to perform functions with government employees or to contract them out to the private sector (Government, 1990).

Provider Sophistication Versus User Simplicity: European Servicing Through Bridging the Gap

by Per Nielsen [1]
Danish Data Archives
Odense, Denmark

Introduction: Working on the right problem
When I first entered the "data archive movement" (February 1, 1974), everybody seemed to be very preoccupied with the creation of advanced software for the mainframe: Report generators (even though there was litte to report on), search systems (even though there were few surveys to search among), and data base systems. Later on, we realized that the *de facto* standards were developed at larger organizations - either within (OSIRIS from the ICPSR, SPSS from NORC) or outside (SPSS Inc., SAS in Raleigh, and all the other business firms all over the marketplace) "our world" of social science research institutes and data archives. I am sure that we spent quite a lot of time during the early years working on the wrong problems given the needs of the time; however, the work initiated the habit of engaging in cooperative projects among the European archives. This good working habit has survived ever since; as institutions and as individuals alike, the European archives have a close and pleasant collaboration program, ranging from responding to incoming servicing requests over staff-relevant Expert Seminars[2] to Business Meetings once or twice a year.

One topic that was hardly ever discussed during these early years (even internally among data archivists) was the actual level of servicing provided by the data archives during a given year; it was the tacit understanding that the actual (quantitative) level of servicing was not a topic that would underline the *raison d'être* of the new data organizations, data archives, and data libraries.

Now, almost twenty years later, we have vast amounts of data sets in custody, and the demand for data for secondary analysis has increased dramatically, not least with the advent of the PC. Unfortunately, we are not so well-prepared to meet this demand as one would expect, given the advanced techniques developed and applied in the take-off phase. In Babel-like Europe, there are still huge obstacles to a free data flow from data providers to data users. We are working on these, but there is a long way to go. At a recent seminar, two American scholars[3] told us (the data archive professionals) that we had been overtaken by the "ordinary" library people in terms of computer mediated communication. Shame on us!

The real problem during the sixties and seventies, looking in the rear-view mirror, was to localize and collect data, to develop standards for documentation, to teach data collectors "sound methodological/technical practices" during the data generating process, and to store the data safely - with a long term archival perspective in mind. We did perform all of these tasks once we found out that we could import most of the software tools from the outside; however, then we worked little on the search systems and the other advanced tools that successively became relevant as we had thousands of data sets in our holdings; in many archives, the then advanced retrieval systems of the seventies were maintained and slightly developed; some of them are still in use in the early nineties.

The "right problem" is simple: To find relevant data sets that meet the specifications of a user and transfer the data to that user - with the shortest possible elapse time and with the least possible input of human and machine resources in both ends of the communication line.

The topic: *Remote Access and New User Services*
My impression is that American scholars find it very difficult to get an overview of the European data marketplace; honestly, it is sometimes difficult even to Europeans working in the market!

In this paper, *Remote Access* is interpreted as "Access from North America to European Data" - rather than looking at the technical notion of remote access (i.e. running jobs on a distant machine). The underlying philosophy is that the user is substantively oriented rather than technically fascinated; the user would rather have data available in a known environment than shuffle around in dozens of differently functioning systems to dig out what (s)he needs.

Given this interpretation, *New User Services* will, to a certain extent, become equivalent with present user services. Some 80% or more of the DDA-servicing is domestic (i.e. national), so the new services will be developed for the national market first. Consequently, I can claim to be knowledgable about the Danish situation, only; and, being a native of a small country with a peculiar language, I realize that this situation may be of

minor interest in North America.

Finally, my personal feeling is that a paper on *Remote Access and New User Services in Europe* would be much better in 1995 (third year of the Open Internal Market, cf. below); after a couple of decades with consolidation on the (national) *archival* side of the data organizations, we shall now move into an era where the (national and international) *servicing* aspects gain more weight. This gradual shift in emphasis is (among many other indicators) reflected by the fact that the ECPR Council has accepted the theme of *Integrating the European Data Base* for the ECPR Joint Sessions of Workshops 1992 (Limerick, Ireland)[4].

The scene: The integrated Europe (United States of Europe?)

Being in North America, a notion about the EEC-generated phantom of European Integration is perhaps necessary. With the introduction of the Open Internal Market (end of 1992) and the current plans regarding an economic and monetary union (three stages during the nineties), with or without a political (and maybe even foreign policy and defence) union, the Brussels establishment (especially the Commission) has succeeded in moving some frontiers of thinking, especially in business - and maybe even more so in North America and Japan than in Europe!

Even though the United States of Europe is being discussed, by supporters and adversaries alike, this vision will remain a phantom to most Europeans for another couple of decades. In the perspective of *Integrating the European Data Base*, the differences of language, ethnicity, culture, religion, wealth and political and social science tradition represent obstacles to the free data flow; the same is true of the differences in economic development and technological sophistication - a gap which is more evident in the East-West dimension than on the North-South continuum.

Furthermore, it is important to notice that the cooperation between European research institutions (and hence social science data archives) has never been limited to European Community members; it has been open to any institution with the interest and the capacity to participate. Generally (and this is especially true with respect to the data archives) the nation-state has been the represented unit. It has been difficult in some countries to find the relevant (i.e. nationally representative) institution; and this problem area will return to the scene as new states emerge (East) and as some countries develop specific institutions to deal with data archiving (e.g. specialized historical data archives in the larger countries).

Finally, with a landscape of "peer partners" among European archives, it has not (yet?) been possible to establish the *European Data Archive* [5] that might be the gross dealing agent in Europe, comparable with the ICPSR in North America. Needless to say, it would be easier from the outside to address one central agent that collected and disseminated all major European data sets of broader interest. Our response so far to this demand is: Contact any one of the archives, and they will (ideally!) let the message pass to everybody else. As a matter of fact, this procedure has proven to be efficient on a number of occasions; but please be accurate and specific when elaborating the request!

Which are the heavy resource demanding servicing tasks?

In the American context, the European archives should be understood as an amalgamation of the gross dealing agents (e.g. the ICPSR) and the local retail servicing facility (university data libraries, state data archives, etc.) Covering the whole set of archival as well as servicing procedures, we have a good overview of the costs involved in different parts of the whole process.

At the archiving end, it is the cumbersome data processing to a *standard archival format* that digests heavy resources. Even though standards may vary from place to place, most of us want to produce standard codebooks (in a format derived from the OSIRIS dictionary-codebook format - type 3 to have a clean ASCII character set.) Most of us want to have a good study description (sub-structured in a more detailed way than the ICPSR free text study description), and most of us want the data to be immediately accessible for the major analysis packages like SAS and SPSS.

At the servicing end, the gross-dealer functions take little time: Users requesting specific data sets that have already been processed to the standard archival format can be serviced from one day to the other - or even within hours; they will receive easily accessible data and can start off with their analyses immediately, cf. below. The resource-consuming customers are those that want to obtain data that meet certain search criteria (often too vaguely defined!) - and this is especially time-consuming if the user wants to perform cross-national comparisons and/or if non-standard data sets are involved.

Obstacles facing the user of European data

Below is a list of existing obstacles that the data user may face when trying to get hold of data relevant for (cross-national) analysis. Imagine that the user asks the local national archive in one country to facilitate access to data from several European countries, relevant to a certain topic; what is the process ahead?

1. The local archive sends out a "search warrant" to all other European archives. This can be done quite quickly via E-mail. However, most European countries still do not have a data archive; and some of the existing national archives are heavily under-staffed, so that the requesting archive will get a late reply or even no response at all. Result: With good luck, relevant data will be found in 3-5 countries, only.

2. Some of the data sets localized may have been produced by central statistical offices (CSOs) or administrative agents; in these cases, data may not be available at all - or the user will have to go to the country in question because data export is not allowed. Another possible obstacle is an embargo period for secondary analysis, introduced by the primary investigator.

3. Some of the relevant data sets may have an inappropriate format, i.e. they are not immediately available for analysis with SAS or SPSS. It may take months or even years until the data archive has improved the technical availability of the data set.

4. The data user may have to sign undertakings with each of the primary investigators before the data can be delivered for secondary analyses.

5. The documentation of the relevant data sets may be available in the local language, only; and this, of course, is the rule rather than the exception.

6. Remedying obstacles 1-5 may cost real money that the user may not have available.

7. The (few) data sets that actually pass the obstacles 1- 6 will now be sent to the requesting archive; they in turn pass the data sets on to the user.

8. If, during the secondary analysis process, problems arise with the data or the documentation, the trouble-shooting will be quite difficult also.

The above picture of the obstacles is quite pessimistic, some European data archive people might say; unfortunately, my feeling is that it is a realistic one. There is a long way to go for the archives united in CESSDA[6] (Committee of European Social Science Data Archives) before we have an Integrated European Data Base.

With 15 years of fieldwork, we have not yet fully implemented the visions presented in a paper at the CESSDA founding meeting by one of the fathers of the Data Archive Movement, the late professor Stein Rokkan: "Our basic philosophy is very simple: we do not believe the archival movement in Europe will get anywhere unless there is a real break with the tradition that archives are there simply to store, clean and reformat separate data sets. The future lies with active reorganization of data: linkage across files, build-up of time series sets, preparation of handy packages for use in the classroom, integration of packages with better computer routines for graphic display, cartography, visual model-to-data fitting." (Rokkan's underlinings).

Remedying the obstacles: Remedying the obstacles: State-of-the-art and planned activities

We are doing our best to try to smoothe the facilitation of European data to the user. Let's reiterate on the obstacles mentioned above and see what is being done and what can be done within each of the "obstacle fields" identified in that section. Doing so, we shall look at the European level first, and I shall add a few comments about the Danish situation - which I know best!

Localization of relevant materials

Most archives do have printed *catalogues of holdings, with multiple indexes*, from which you can figure out whether the other archives do have relevant data sets in a given field. However, the printing is expensive, and the paper-bound inventories tend to become outdated quite quickly - both with the acquisition of new data and in terms of "processing classification," access restrictions, etc. It is possible, of course, to acquire the machine readable text from the catalogues of each archive and search these in your own retrieval environment; but this does not solve the updating problem. Consequently, the most evident solution is either to integrate the primary cataloguing at one central location or to search in the catalogues of the other archives, via telecommunication, at their own computer installation.

The first path, *integrating the catalogues*, has been worked on with some energy. The Commission of the European Communities (CEC) had actually granted money that would allow catalogue integration in a project with DDA, ESRC-DA, STAR and ZA as the major project partners. This project stranded because the CEC demanded that the resulting data base had to be commercially viable after the 2-year project period. (It probably would not have been after 10 years; but the CEC bureaucrats, preoccupied with "commercialization," are not at all sensible to the special problems in the academic sector!)

The second path, *searching via telcommunications*, is

probably a more realistic one. Given that most archives now have TCP/IP and FTP facilities available at the installation where the catalogue information is stored, the searching as well as the actual exchange of data may take place using these communications and file transfer facilities. This procedure assures the user that the most updated version of the catalogues is searched and the most processed data set is transferred.

Even though many of the archives have retrieval facilities that are open to the user, most searching is still done by the staff of the archives on behalf of the user; this is probably going to be the case in the foreseeable future for all other than very heavy users of data.

In the case of the DDA, we have one central retrieval system, DDAGUIDE, based on the study descriptions. It is available to most potential Danish customers, located at UNI*C (a national computing center for research and education). However, it is not yet open to users without an account number at UNI*C. In addition to DDA-GUIDE, we have several in-house search systems (some mainframe-based, others PC-based) searching the contents of the machine readable codebooks. In order to integrate retrieval at the study description and codebook levels, we have designed an integrated system and applied for money from the SSRC to have the UNI*C-people implement the system; hopefully, there will be a remote user access to this integrated retrieval system[7]. (On the other hand, the language stored will be Danish, cf. below, where the DDAGUIDE stores English language texts!)

Handling access restrictions on the data

In most European countries, the access to process-produced data (administrative and statistical data) is hampered - mainly due to three factors: (1) The bureaucratic traditions of government and a lack of Freedom-of-Information-tradition; (2) the privacy legislation; and (3) the wave of cutting in public spending - indicating that the statistical bureaux (CSOs) and other data owners want to sell their data rather than offer the data for free (or at a quite low price) via the social science data archives. Let us take a look at a few examples:

In *Norway*, where the relations between the central statistical bureau (CSO) and the NSD has been better than in most countries, the NSD can disseminate a lot of statistical data for research and educational uses; this has mainly been done with regional data, where the NSD probably has the largest collection of commune-based data in Europe. On the other hand, there are severe restrictions to the access to survey-data (data on individuals); for instance, the later election studies in Norway have been collected by the CSO; consequently, the data may not be taken out of the country, so that the user will have to go to Norway to use such data sets.

In *Sweden*, also, the election study data sets and other CSO data on individuals may not leave the country; the SSD has tried to apply the LIS-model (Luxembourg Income Studies) to gain indirect user access to such data[8].

In the *United Kingdom*, the ESRC-DA is "re-selling" selected data series from statistical authorities to the academic community. Also in *Hungary*, Tárki has quite close relations with the statistical office in Budapest.

Data from the *academic community* are usually more easily available than data from the CSOs. Even so, it is in some countries (for some studies) necessary to ask the primary investigator's permission. It seems to be generally accepted that the primary investigator may impose up to a 2-year embargo on the data.

In *Denmark*, the CSO (Danmarks Statistik) tries to make a lot of time series data banks available on a commercial basis. Four rather large "data banks" are available, covering national economic time series (DSTB), commune statistics on the 275 local administration units (KSDB), labor market statistics (ABBA), and business related statistics (ESDB). However, the Danish CSO is very reluctant to release survey data of any kind. The big public survey organisation, The Danish National Institute of Social Research, on the other hand, generously puts all its surveys (in a *de facto anonymous* form) at the disposal of the DDA and her users - free of cost. This fact demonstrates that it is the *interpretation* of the data legislation rather than the Acts themselves that render access impossible.

Handling technical problems with the data

There is little standardization across Europe with respect to the *processing classes* (cf. ICPSR's Classes I-IV - a data class structure which is presently being revised). Some archives (e.g. DDA, SSD, ZA) follow a strategy pretty much like that of the ICPSR, having machine readable codebooks for their top-class studies. Other archives (e.g. NSD, STAR) try to make as much as possible available as SPSS Export files - not necessarily having all information from the questionnaire (or other instrument) in the machine readable codebook. The ESRC-DA has realized that they do not have the resources to process their several thousand studies to the "ICPSR Class I"-level; instead, they have developed a thesaurus and apply that to append relevant search entries to the study descriptions in order to have a search base without having fully-fledged codebooks for all studies.

In general, most European archives aim at making the data sets available for analysis with SPSS and SAS.

However, most archives hold many data sets that have not (yet) reached this level of processing (ICPSR Classes II-IV); for some of these data sets, the user will have to produce the setup on his own, based on a card-image data set and a paper documentation thereof.

At the DDA, we stick to the traditional "ICPSR-type" of documentation with respect to the codebook, adding a (structured) standard study description to this codebook. Streamlining the data documentation and processing work (mainly based on programs such as SAS, KEDIT and REXX in an OS/2 networking environment) we can more than keep pace with acquisitions, so that the number of "non-Class I" data sets is diminishing.

Data ownership as a restriction to remote access

Whereas many archives have *searching facilities* available for remote access from the users, few archives have the data files as such available for immediate analysis. This is usually not due to technical restrictions; rather, it is based on proprietory considerations: The principal investigator is the official owner of the data, and sometimes her or his written consent is required in each individual case before the archive can offer access to the data set itself.

This type of restriction should problably be removed in the future; given the fast technological development (with communications protocols like TCP/IP and file transfer protocols like FTP on the one hand, and distribution on CD-ROM or other mass storage media on the other hand), the provision of free distribution should be granted to the data archives by primary investigators.

Vis-à-vis the researchers, the DDA has not yet found the formula that will allow access to all stored materials without prior written consent from the depositor. However, access is never denied, so we consider it feasible to get an agreement about unrestricted access with most donors, once we really need that - either in order to allow remote access for analysis on our computing facilities or to distribute selected data sets on mass storage devices.

The tremendous language problem in Europe

Within each country, it is considered "normal" or even indispensable that the documentation be produced in the national language; the only general exception to this rule seems to be The Netherlands, where Steinmetzarchief produces SPSS-setups for all files in English rather than in Dutch. Most archives do have catalogue information available in the English language, but the codebooks and/or the questionnaires (or other instruments of data collection) are available in the national language, only. This is an obstacle to remote (*in casu* foreign) access - to which there is no readily available solution.

Everybody who has been engaged in cross-national research projects will know that it is a tremendous problem to produce cross-culturally comparable data, in part due to the language problem. This is true even in the culturally relatively homogeneous European Community (reflected in the Euro-Barometer surveys); but the difficulties are even greater if one goes to Second or even Third World nations (which, for instance, the ISSP-program is doing).

Definitely, there is not enough resources within the European archives to produce all documentation in the national language *and* in a world language (e.g. English). All users of European data should be aware, consequently, that they will have to be able to read the language of the nation under investigation - with the aforementioned exception of The Netherlands. (Some scholars would argue that you would have to know some language and culture prior to engaging in quantitative (or qualitative) investigations of a specific nation *anyway*; we shall not engage ourselves in that discussion here.)

At the DDA, we keep study descriptions in both Danish and English; we can, therefore, inform about our data (in catalogues and *ad hoc* listings of selected topics) in either language. But with the codebooks it is different; even though we have produced some English language codebooks (in addition to the Danish ones) for a few frequently exported data (e.g. the *Continuity Guide to Election Studies*, which is also disseminated through ICPSR), the bulk of the codebooks are available in the Danish language, only.

It may be a future project among European archives to produce English language codebooks within areas where cross-nationally comparable data sets can be "constructed" by the archives. Such projects have been successfully carried out in the past; for instance, at lot of National Election Studies (and Continuity Guides based on these) have been produced in cooperation with the ICPSR and are now distributed from Ann Arbor to the membership.

Fee schedules among the European archives

In some countries, the servicing of users is free of cost (apart from "media" such as paper-codebooks, diskettes, etc.); other archives have a fee schedule - the size of the fee depending on factors such as size and complexity of the data sets delivered, staff time involved, etc. Sometimes there is a discriminatory pay-schedule, where students are at the cheap end and business applications in the expensive end - somtimes with researchers in a middle position. It is beyond the scope of this paper to try to spell out all the fee schedules; they change now and again, so the user has to ask in each specific case.

At the DDA, servicing of archived files is free (media cost recovery is demanded if the media are not returned). Also, the staff and machine time consumed by performing searches for users is free. However, special services such as "super-quick processing" of DDA-studies, processing of requestor's own data sets, or translation of codebooks will have to be paid for by the requestor.

Actual transmission of data to the user

As a consequence of an agreement between all CESSDA archives, there are certain rules that apply to international data transfers; the major contents of this agreement can be summed up as follows:

* Each national archive is the primary repository regarding data from that nation. ("Fishing-zone agreement").

* All requests from within one of the "CESSDA-countries" should be directed through the home archive. (This seemingly bureaucratic rule is administered liberally: If a foreigner requests Danish data, for instance, we would always inform the relevant "home archive" and, if they so wished, send the data via that archive). The idea is, of course, that the users should not be able to circumvent pay schedules by going abroad to the free or low-cost archives.

* Cross-national data sets are processed and disseminated according to mutual agreements among the relevant archives.

As mentioned before, the actual transfer of an already processed data set is not very time consuming. However, the actual procedures may differ from place to place; needless to say, in inter-archival transfers the technicalities would be agreed upon aforehand (if they are not already known from earlier transfers).

At the DDA, the transfer is done according to the specifications wanted by the actual user. If the user works on a mainframe, we would normally send the data to that mainframe from our central archive (magnetic tapes at a central UNI*C mainframe). If the user wants to work on a PC - which some 90% of users do, we will send the data on a diskette, containing a DOS BAT-file that will do all the work necessary before the analysis: Make the necessary directories on the user's hard disk, copy the files onto the hard disk, UNZIP the packed files, and, if necessary because of multi-volume delivery, put split files back together.

Let us assume that the study number DDA-9999 was requested and sent on one or several diskettes; after the user has run a DDA-COPY.BAT job, (s)he now has the following files in a directory with the name C:\DDA-9999:

DICB9999.OSI	**(OSIRIS-like dictionary-codebook, ASCII)**
DATA9999.OSI	**(OSIRIS-like data file, ASCII)**
OSI-SPC.EXE	**(DOS-program for system file, cf. below)**
LIST9999.PRT	**(ASCII-listing file with SD and codebook)**

All the user has to do now is to run the OSI-SPC program (a DDA-utility) which will ask (1) whether the user wants an SPSS or a SAS file; (2) which variables the user wants to include in the systems file; and (3) the names of the dictionary-codebook input file and the setup output file. When OSI-SPC has finished (in seconds, even with very large files), a setup is ready to build the specified SPSS or SAS systems file - with all necessary variable and value labels in place. Again, the user needs to specify only the data set names before running the systems file generating job. The LIST9999.PRT file is just a stream of lines that can be printed on any type of printer; the idea is that the user can save the money for the printed documentation if (s)he prefers to run off a printed copy instead.

The supply of data on diskettes takes place from a copy of the original archive which contains the ZIP-files; this "archive copy" is kept on a 1 Gb traditional disk on a net server, so there is very quick access. Needless to say, such files can be sent over the external networks to the user instead of using diskettes; however, the DDA is waiting for the OS/2 version of TCP/IP and FTP - which is to be in the market very soon.

Our experience is that users (and we ourselves, when receiving foreign data like that) are very satisfied with the present procedure.

Problems with data or documentation during secondary analyses

It is important to know which archive is responsible for the data sets that "drift around" in the international social science community; otherwise, when errors or omissions occur during the secondary analyses, it is close to impossible to remedy or clarify such problems. A couple of examples will demonstrate this problem:

parties (Venstre, i.e. the (Conservative) Liberals) are absent from data as well as from documentation. Who "produced" the error: The principal investigators (Jacques-René Rabier, Helene Riffault, Ronald Inglehart); the Danish data collector (Danish Gallup); the international coordinator (Faits et Opinions); the Zentralarchiv (where the file is first available); or the ICPSR (where the final documentation is produced)? - Needless to say, the error is critical to a political scientist who will use this variable extensively during the analyses.

A Danish user participating in the repeated International Value Project finds out that there are problems with the "oversampling" of young people in the 1981 Value Project file for Denmark. There were no Danish researchers involved in the 1981 Value Project; all you can do is to ask the Danish collector (Observa, which has changed name and staff in the meantime); the Coordinator (British Gallup); or the involved archives (*in casu* the ESRC-DA, but we got the file from NSD). It is more than likely that you can never solve the problem! (Which, in this case, derives from the fact that archives entered into the process of preserving the data long after the primary investigation.)

In conclusion: What is achieved, and where are we heading?

In the European countries with a national data archive, a huge data resource is immediately available for analysis to everybody who knows the human language of the country; the range of users has been augmented laterally (data in schools[9] and easily operated analysis packages[10]) and horizontally (the concept of "social sciences" is broadening with "new" disciplines all the time.)

The users want the data "packaged" to their needs, on their own computer and for their own analysis package; it is my feeling that we have accomplished this task on most levels of user sophistication: The number of users is constantly rising, and the servicing is becoming very efficient in the archives. To move from diskettes to FTP transfer of data is a technical detail of minor importance to the user; however, it may save resources among the data suppliers and make the transfer across long distances quicker.

The broadening of the topical area covered by the archives is a process that may take different paths in different countries. Take the historical data as an example: In Germany, the Zentrum für historische Sozialforschung started its careeer as an independent institute; later, the ZHSF was moved to form a department at the ZA. In The Netherlands, a historical archive is on the steps; right now, however, funding is lacking. In the UK, the situation is under review.

In smaller countries, it seems likely that the existing archives will cover the "new" disciplines. In Denmark, data from history as well data from social medicine are archived at the DDA - and have been for years. The DDA hosts the up-coming AHC Conference 1991 to demonstrate that fact to everybody in Europe[11].

Obviously, the immediate future in Europe will be devoted to the project of *Integrating the European Data Base*. CESSDA is right now "incorporating" (with a formal Constitution and membership fees) as one step in that direction. Projects are underway that will facilitate searching across archives. Data exchange will take even less time in the future using telephone lines for transmission rather than snail-mail with data media.

Let the remote users come; we shall give them access and demonstrate that our services are better and quicker than ever before!

Footnotes:

[1] Presented at the IASSIST 91 Conference held in Edmonton, Alberta, Canada. May 14 - 17, 1991.
The Danish Data Archives (DDA) is a national social science data archive established in 1973. Since 1978 the DDA has been located at Odense University; some time in 1991, the DDA is likely to be relocated, most likely to be a department of the Danish National Archives. Danish Data Archives, Munkebjergv{nget 48, DK-5230 Odense, Denmark, E-mail: DDAPN@NEUVM1.BITNET or ddapn@vm.uni-c.dk, Telephone: (+45) 66 15 79 20 x 2810, Fax: (+45) 66 15 83 20

[2] The European archives arrange so-called Expert Seminars for staff-members, hosted by one of the archives, once or twice a year. It should be noticed that the data archive "milieu" is more institution-based than person-based compared to the situation in North America, cp. the large IASSIST constituency in North America compared to Europe.

[3] Professors Harold Clarke and Mark Franklin, Texas, made these comments during the ECPR planning session *Integrating the European Data Base* (Essex, UK, March 22-26, 1991).

[4] This Workshop, co-chaired by Ekkehard Mochmann (ZA) and Eric Tanenbaum (ESRC-DA) was being prepared in March of this year, cf. note 3 above.

[5] Unfortunately, a local research institute in Mannheim (FRG) has adopted the name EDA (European Data Archive); this may confuse some users inside and outside Europe. The name is misleading also in the sense that the holdings of EDA are mostly "second hand data" (i.e. data from other archives).

[6] CESSDA was founded in 1976 in Amsterdam (May 31 - June 1) as an informal cooperation between existing social science data archives. CESSDA is the European branch of IFDO (International Federation of Data Organizations). Active European archives are - with an asterisk (*) in front of the CESSDA founding members:

ADB (All-Union Data Bank), Moscow, USSR
* ADPSS (Archivio Dati e Programmi per la Scienze Sociali), Milan, Italy
* BASS (Belgian Archives for the Social Sciences), Louvain-la-Neuve, Belgium
BDSP (Banque de Données Socio-Politiques), Grenoble, France
* DDA (Dansk Data Arkiv), Odense, Denmark
* ESRC-DA (Economic and Social Research Council Data Archive), Essex, UK
* NSD (Norsk Samfunnsvitenskapelig Datatjeneste), Bergen, Norway SSD (Svensk Samhällsvetenskaplig Datatjänst), Gothenburg, Sweden
* STAR (Steinmetzarchief), Amsterdam, The Netherlands
TARKI (Social Science Information Center), Budapest, Hungary
* ZA (Zentralarchiv fuer empirische Sozialforschung), Cologne, FRG
WISDOM (Wiener Institut fuer sozialwissenschaftliche Dokumentation und Methodik, Vienna, Austria

Other countries, e.g. Czekoslovakia, Ireland, and Switzerland, are expected to set up similar archives in the near future.

[7] This system represents the first stages of a larger project presented by Karsten Boye Rasmussen at the IASSIST Annual Conference 1990. The remaining parts, having searched study descriptions and codebooks, include an automatic downloading of the relevant data sets for analyses.

[8] The LIS-model aims at securing access to confidential data in the following manner: Instead of distributing the "real" data, a constructed "model data set" with the same distributional characteristics is disseminated. Once the user has produced the setup that generates the right analyses, (s)he sends that setup to the data base administrator - who then runs it against the real data. The administrator sends the output to the user after checking that no confidential information can be disclosed in the output. (Setup and output can be sent via electronic mail.)

[9] Most "CESSDA countries" have made teaching packages based on some of their more "popular" data sets for undergraduate or even school students. In Denmark, some 25% of all schools acquired a teaching package during the second half of the eighties.

[10] NSDstat from the NSD is presently being distributed in several of the other "CESSDA countries". NSDstat was presentedand demonstrated during one of the workshops prior to theIASSIST Edmonton Conference.

[11] Association for History and Computing 6th International Conference will take place in Odense, Denmark, August 28-30, 1991. Usually, several hundred historians from all over Europe (and some overseas guests) participate in the AHC Annual Conferences.

The Depository Distribution of CD-ROMs: A Review of the First Year

by Juri Stratford [1]
University of California, Davis

Introduction
A large part of the work of depository librarians is providing public access to the vast number of statistical publications produced by various agencies in the Federal government including most notably the Census Bureau and the Bureau of Labor Statistics. Until 1989, the distribution of Federal statistical data in machine-readable form was limited to programs administered by individual agencies, e.g. the Census Bureau's State data center program and the Bureau of Economic Analysis' local area data program, and institutions acquiring data directly from agencies for purchase or through consortiums and archives such as ICPSR.

The first CD-ROM to be distributed to depository libraries was **Test Disc 2. Test Disc 2** included state and county data from the 1982 Census of Agriculture and zip code data from the 1982 Census of Retail Trade. This was distributed to a few libraries on an experimental basis in 1988 and was made available through regular depository distribution in 1989 shortly followed by distribution of the **City and County Data Book** CD-ROM in 1990.

Since then the Census Bureau has distributed a number of CD-ROM products through the depository system including data from the the 1987 Economic Censuses, the 1987 Census of Agriculture, and the reapportionment data from the 1990 Census of Population and Housing. Other significant CD-ROM products distributed to depositories include the **National Health Interview Survey** produced by the National Center on Health Statistics, the **Toxic Release Inventory** produced by the Environmental Protection Agency, and the **National Trade Data Bank** produced by the Commerce Department. A number of other agencies including the Department of Defense, NOAA, and the Geological Survey are also beginning to distribute CD-ROMs through the depository system.

CD-ROM and Papercopy Distribution
In some cases, the depository distribution of CD-ROMs complements the depository distribution of paper or microfiche products. For example, the EPA's **Toxic Release Inventory** was made available simultaneously to depositories on CD-ROM and on microfiche. The CD-ROM distribution of the 1987 Census of Agriculture and the 1987 Economic Censuses followed the distribution of these publications to depositories in paper copy. This has also been the case, so far, with the CD-ROM distribution of the **City and County Data Book** and **County Business Patterns.**

In other instances, the CD-ROM distribution provides depositories with materials that they might not have had otherwise. For example, the CD-ROM distribution of the **National Health Interview Survey** data, the PL94 data from the 1990 Census of Population and Housing, and the zip code data from the 1982 and 1987 Economic Censuses represent data not available in paper copy.

A final area, which should concern the depository community, is the replacement of the depository distribution of paper copy or microfiche with CD-ROM. For example, the monthly import and export data from the Census Bureau is now only being distributed to depositories on CD-ROM. Also, the 1985 Congressional Record CD-ROM was recently distributed to depositories on a trial basis as a replacement for the bound edition.

Advantages and Disadvantages of Data on CD-ROM
CD-ROMs offer some advantages both to end users and data producers. The electronic distribution of data provides the potential to enhance user access. Libraries can provide access to a vast quantity of Federal data in machine readable form allowing end users the ability to work with the data on their own microcomputers. Where in the past researchers might have had to extract data subsets from tape, or to key in data by hand from printed sources, they can now copy the files directly from the CD-ROMs to diskette. Data from these CD-ROMs are available free of charge and without copyright restrictions.

CD-ROM also offers some advantages to data producers. In many instances, it is less expensive to produce and distribute CD-ROMs than paper copy, and Congress is anticipating a cost-saving through CD-ROM. In a statement before the American Library Association Legislation Committee in January, Robert W. Houk, the Public Printer reported that GPO requested a fifteen percent increase for salaries and expenses primarily associated with the distribution of 1990 Census publications. However, Congress reduced the request from

$27.9 million to $26.5 million, projecting that the Census Bureau would distribute a greater proportion of documents in CD-ROM formats, rather than in microfiche as originally anticipated[2].

When materials in paper copy or microfiche are replaced by CD-ROM, the CD-ROM distribution may be viewed as shifting the expense of production from the data producer to the data user. While many depository institutions are facing budget reductions, they are now also faced with the added expense of devoting CD-ROM stations to provide public access to the Federal data. Other expenses will include acquiring the proper software to work with the data and providing paper for printouts.

Data Format and Software Issues

The Census Bureau data are being distributed in dBase format. For the demographic data, the Census Bureau is distributing software to display tables, usually for specific geographic areas, to the screen or to a printer. The software for the foreign trade data displays data for the current month and beginning of current year to date for the most specific commodity code. The Census Bureau is also distributing programs such as Extract to create small data subsets for output as ASCII files, dBase files, or Lotus worksheet files. Extract can also print tables from the data sets.

The Extract program requires a large number of dictionary files describing the data. The dictionary files for the Economic Census are about four megabytes and these files have been included on the CD-ROMs. However, in other instances these files have been distributed separately on floppy disks, and they must be installed on a hard disk for use with the CD-ROMs. The **City and County Data Book** files require 640kb of hard disk space, **County Business Patterns** files require 650kb, and the monthly import and export files require 3.2 megabytes.

As the Census CD-ROM files are in dBase III format, the data can also be accessed directly using third party database management and spreadsheet software. However, many of the Census Bureau files are large and are very difficult to work with on a microcomputer. The **City and County Data Book** files are small enough, and most of the Economic Census files are less than one megabyte, though a few are as large as two megabytes. However, some of the PL94 files are as large as 300 megabytes; and the foreign trade export files are currently about 258 megabytes while the import files are more than 550 megabytes.

Public Access Issues

Regardless of what software is used, whether a library uses Extract or dBase or some other software, data extraction from these files requires a lot of processing time. It appears unlikely that Extract or dBase could be used to access these data files in a public reference area. In a recent article in **Government Publications Review**, Steven Staninger documents the time required to conduct a few simple searches using **dBase III Plus** with the 1982 Census of Retail Trade data on Census Bureau's **Test Disc 2**. In one example, it required twelve minutes to execute a search for a single type of business code in the California data file on an IBM PC[3].

There appear to be two strategies to the problem of public access to numeric data files on CD-ROM. A combination of both tactics will probably be necessary to provide adequate public access.

The first strategy is to develop the personnel and physical resources necessary to effectively deal with the problem of data extraction. To adequately deal with electronic formats a depository will require at least one staff member with microcomputer expertise and a strong social sciences background. At a minimum, this person will have to be familiar with DOS, database management software and spreadsheets; a familiarity with statistical programs such as SAS and/or SPSS would also be helpful. This person must also be familiar with the printed sources and have a close working relationship with the local data archives facility, if any, in order to know when CD-ROM is more appropriate than paper copy or data on tape.

There must be adequate physical resources as well. There appears to be a well established base of CD-ROM stations in large depository libraries. This is supported by the fact that at least half of all depository libraries have selected some Census data on CD-ROM*. Many of these depositories are also using these CD-ROM stations to provide end-user access to bibliographic files. However, as Staninger's article suggests, the effective use of the Census Bureau CD-ROMs will not lend itself to this setting. Libraries will need to provide CD-ROM stations out of the reference area where either library staff or end-users can extract data from the CD-ROMs. These CD-ROM stations will also need to devote a large amount of hard disk space to work with the depository CD-ROMs. In addition to the requirements outlined above, many of the depository CD-ROM products have their own front ends requiring a large amount of dedicated hard disk space, e.g. the **National Health Interview Survey** requires about five megabytes and the **Toxic Release Inventory** requires about seven megabytes of hard disk space.

The second strategy is to look for solutions in the private sector. While there are a few products, such as PC Stars, which are designed to work with the depository CD-ROMs, many commercial software producers are

reselling the Federal data. For example, Space-Time Research's **Supermap**, marketed in the U.S. by Chadwyck-Healey, contains data from the 1980 Census of Population and Housing Summary Tape Files, STF1-C and STF3-C, and additional county and land area data; and **StatMaster** produced by CyberSoft includes data from the **County and City Data Book**. While these commercial products might be expensive, each of these products provides enhanced access to the data.

There appears to be a fear that this could make depository libraries more dependent upon the private sector for access to Census materials; a Census Bureau report notes that librarians are "concerned by the need for user-friendly software to access census data and fear that it may be available from only the private sector at prohibitive cost."[5] But there is a long established, successful history of private publishers providing bibliographic access to depositories. As Sir Charles Chadwyck-Healey stated in a recent interview published in **Government Publications Review**: "Governments seem to be extraordinarily bad at distributing information efficiently. It is probably inevitable. I am not sure there is going to be an enormous advantage to having very cheap data available from government sources if those sources are not able to disseminate it in and efficient and effective way."[6]

The depository CD-ROMs will have a great impact on library public services. In fact, if widely adopted, the CD-ROM distribution has the potential to transform the typical depository as we know it. Perhaps the greatest impact of CD-ROM in a library is the increase in workload for the public services staff in whose area the CD-ROM is located.

In a recent article, Steven Zink argues that the positive aspects of CD-ROM have tended to overshadow the human resources required for its use. He notes that "a persistent administrative malady is the assumption that technology will decrease, or at least not require additional, demands on staff time." In fact, the opposite tends to be true. Zink explains that "while technological advances may have resulted in personnel reductions in selected technical service areas, the use of automation where the public directly confronts technology has generally increased the need for user assistance."[7]

Depositories have yet to determine how best to deliver the additional assistance that users of electronic formats will require. Generally, document librarians perceive that the data extraction from the depository CD-ROMs will follow the patterns established by mediated online searches. However, as Elizabeth Stephenson observes: while many librarians are experienced in handling bibliographic data on CD-ROM, "few have any experience or training in the manipulation of numeric files." She believes that librarians will have to become familiar with the hierarchical structure of the files and statistical languages to effectively work with the depository CD-ROM products.[8]

Conclusion

In 1988, Diane Smith of Pennsylvania State University conducted a survey to determine the preparedness of depositories to provide access to electronic data products. She examined the extent to which depositories were experimenting with the provision of electronic services and looked for characteristics common to innovative libraries. Her survey covered plans to include documents in local online public access catalogs; the use of online databases, CD-ROM, statistical software, and expert systems in depositories; and available hardware within reference areas. Smith concluded that depositories were ill-prepared to deal with electronic formats and that there is a definite need for depository libraries to face the training and collection management challenges presented. She cautioned that "if this work is not done, it appears that there will be a major crisis in the ability of libraries to deal with electronic data; a crisis that questions the viability of the present situation."[9]

As early as 1988, Jones and Kinney argued that when document librarians assume the responsibility of retrieving numeric, textual, or bibliographic information from computer tapes, they must either know how to program or work with a colleague who programs[10]. It is unlikely that a majority of depositories will be able to provide adequate programming assistance in-house. This means that many depositories may need to establish collaborative relationships with other units to adequately service the depository CD-ROMs. Likely partners include computer centers, data archives and social science research units or teaching departments.

At present, both product development and participation in the distribution of CD-ROMs through the depository program appears to be coordinated at the agency level. For example, within the Commerce Department, the Census Bureau is distributing CD-ROMs through GPO's depository program while the Patent and Trademark Office is only offering their own CD-ROM product, CASSIS, for sale or through their own patent depository program. Each agency is approaching the data format and user-interface issues differently. While the Census Bureau has committed itself to distribute data to depositories in dBase III format for use with privately-produced software, the Office of the Secretary is distributing the National Trade Data Bank CD-ROM with its own unique user-interface and data format. Some CD-ROM products will require that depository libraries acquire privately-produced software to work with the data; others will require that depository staff invest a substantial amount of time mastering the user-interface provided by the agency. While the true magnitude of this trend has yet to

be determined, the impact upon public service in depository libraries will be siginificant.

[1] Presented at the IASSIST 91 Conference held in Edmonton, Alberta, Canada. May 14 - 17, 1991.

[2] Robert W. Houk, "Remarks before the American Library Association Midwinter Meeting: Legislation Committee, Information Update," **Administrative Notes** 12 (February 22, 1991): 1-6.

[3] Steven W. Staninger, "Using the U.S. Bureau of the Census Test Disc 2: a Note," **Government Publications Review** 18 (March/April 1991): 172-3.

[4] Peter Hernon and Charles R. McClure, "Electronic Census Products and the Depository Library Program: Future Issues and Trends," **Government Information Quarterly** 8 (1991): 61.

[5] Sandra Rowland, "The Role of Intermediaries in the Interpretation and Dissemination of Census Data Now and in the Future," reprinted in **Documents to the People** 18 (June 1990): 81.

[6] Jean Slemmons Stratford, Juri Stratford and Steven Zink, "Applying an "Entrepreneurial Attitude" to the Dissemination of Government Information. An Interview with Sir Charles Chadwyck-Healey, Chairman of the Chadwyck-Healey Publishing Group," **Government Publications Review** 18 (March/April 1991): 134.

[7] Steven Zink, "Planning for the Perils of CD-ROM," **Library Journal** (February 1, 1990): 54.

[8] Elizabeth Stephenson, "Data Archivists: The Intermediaries the Census Bureau Forgot. A Review Essay of "The Role of Intermediaries in the Interpretation and Dissemination of Census Data Now and in the Future," **Government Publications Review** 17 (September/October 1990): 443.

[9] Diane H. Smith, "Depository Libraries in the 1990s: Wither or Wither Depositories?," **Government Publications Review** 17 (July/August 1990): 312.

[10] Ray Jones and Thomas Kinney, "Government Information in Machine-Readable Data Files: Implications for Libraries and Librarians," **Government Publications Review** 15 (January/February 1988): 30.

CD-ROM Publishing: Review, Developments and TrendsCD-

by Paul T. Nicholls & Douglas G. Link[1]
Social Science Computing Laboratory and
School of Library & Information Science
The University of Western Ontario

INTRODUCTION:

Kuhn (1970) argues that major advances in science are not evolutionary, but revolutionary; they involve an unexpected change in perspective. The change does not abandon the previously valid model of research, but it establishes an alternative approach that often yields better results. Academic based CD-ROM publishing has the potential to influence dramatic changes in perspectives of education and research. For example, the University of California at Irvine has produced a CD-ROM disc called **Thesaurus Linguae Graecae**, a database, when complete, that will contain all the Greek literature from Homer in the eighth century B.C. to close to the sixth century A.D. This disc is an integral part of the Ibycus Scholarly Computer (ISC), a tool which is revolutionizing classical studies research. Researchers, archivists and administrators have been assembling databases of encyclopedic proportions for decades, but the technology for cost-effective, do-it-yourself publishing of these enormous research databases have only just begun because of CD-ROM technology.

CD-ROM information publishing has been the primary domain of commercial publishers. It is they who have successfully raised CD-ROM to its present level of market appeal. Products such as The Educational Resources Information Centre (**ERIC**) database, The **New Grolier Electronic Encyclopedia**, **Oxford English Dictionary** and **Compton's MultiMedia Encyclopedia** stimulate the imagination of educators, researchers and students alike. Business, government and libraries have been the first to embrace in-house CD-ROM publishing as a cost-effective alternative for distributing and improving access to specialized textual, numeric and image information databases. The support market for in-house CD-ROM publishing has given rise to companies such as Innotech Inc., Meridian Data Inc., Dataware Technologies Inc., Online Computer Systems Inc., OPTIM Corporation, and Knowledge Access International. These companies specialize in providing professional CD-ROM product development services and the sale of complete turn-key systems for supporting in-house CD-ROM publishing.

CD-ROM technology and publishing software is evolving to a point where an individual can actually design, build and produce a CD-ROM disc with a home computer. At the Sixth International Conference & Exposition on Multimedia and CD-ROM in San Jose, California, Sony and Phillips Corp. announced their "orange book" standard which defines a way for WORM drives to write to CD-ROM format. Users of drives based on this standard will be able to create a CD-ROM with a WORM drive and also play existing commercial CD-ROM discs. JVC Information Products will have available by the fourth quarter of this year, the first 5.25 inch, half-height, write-once CD-ROM drive based on the new "orange book" standard. The new drive is expected to be OEM priced at $1000.

CD-ROM TECHNOLOGY

CD-ROM technology emerged in 1983 as a joint effort of Phillips and Sony, who demonstrated their first CD-ROM drive in 1984. The first commercially available CD-ROM database, Bibliofile, appeared in 1985, and the number of available titles has been increasing exponentially ever since (OPTIM 1990). Carlos Cuadra (1991) has documented the vigorous growth in portable, as opposed to online, databases of all types, including CD-ROM, magnetic tape, Bernoulli cartridge and floppy disk:

> One year ago, online databases outnumbered portable databases by a 7-to-1 ratio. The ratio is now about 3-to-1, and closing fast. These figures do not take into account the growing number of portable databases that are being produced or internal use, rather than for sale commercially. These numbers may be growing even faster than the commercial products...

CD-ROM in particular is responsible for much of this growth, and for several good reasons, not least of which is the medium's prodigious storage capacity in relation to other portable media: "While some high-density magnetic floppy disks hold an impressive one megabyte of data, a similarly sized optical disk usually holds 600 times this amount. (Lawrence 1990)"

A 1990 survey conducted under the auspices of the Canadian Library Association's CD-ROM Interest Group (Fox 1990) disclosed that about a third of Canadian

libraries of all types had already implemented or ordered CD-ROM systems. In the case of academic libraries, this proportion was already 44%. Annual surveys by OCLC in the United States have disclosed substantially higher rates of implementation in that country, approaching 100% in the case of academic libraries. CD-ROM has found many types of applications in Canadian libraries and other organizations (OPTIM 1990):

Memorial University of Newfoundland and the University of Guelph have their entire library catalogue on CD-ROM. Ford New Holland uses CD-ROM for its auto parts catalog. Statistics Canada offers bibliographies, directories and census data on CD-ROM. The Department of Fisheries and Oceans made its internal database of reports on fisheries and aquatic sciences available to the public on CD-ROM.

INDUSTRY GROWTH

The Optical Publishing Association (Columbus OH) estimates that CD-ROM revenue from inhouse and commercial publishing and drive sales to be at least US$571 million in 1989, up 41% from US$406 million in 1988 (CD-ROM 1990). "By 1993, according to market researchers Frost & Sullivan, the combined European market for optical disk drives and the optical media will reach US$900 million, up from US$37 million three years ago. (Lawrence 1990)"

According to information in TFPL Publishing's annual CD-ROM Directory, the number of companies involved with CD-ROM activities has risen from 48 in 1986 to 736 in 1989 and 1,840 in 1990.

AVAILABLE CD-ROM TITLES

A survey of commercially available CD-ROM titles was conducted in mid-1990 based on the major printed directories to the medium and using comparative data from three previous annual studies (Nicholls 1991). As of mid-1990, 1,025 commercial CD-ROM titles were identified. At the growth rate that has prevailed over the past few years, well over 2,000 titles are likely available at this time.

Almost half of the titles identified were source databases (containing full text, numeric data, computer software, images or similar data) with indexes/abstracts and directory-type databases accounting almost equally for the other half. The overall proportion of indexes/abstracts on CD-ROM has been declining steadily since 1987, while the proportion of source databases has been rising steadily.

The general/humanities, social science and natural science subject areas are represented almost equally on CD-ROM. Social sciences actually have a somewhat greater share, due to the business and legal databases that (along with medicine) account for 30% of all CD-ROM titles.

The majority of CD-ROMs (63%) are updated annually or even less frequently. The relative proportion of frequently updated titles, (quarterly, for example) has been declining steadily since 1987. This trend is related to the rise in the numbers of source databases, which require less frequent updates than index or abstract databases.

Of all CD-ROM titles, 93% will run on an IBM PC/XT/AT/PS2 or compatible system, while only 13% will run on the Macintosh. Actually, 6% of the titles were designed to run on both platforms, and this proportion is now rising very rapidly. Similarly, although only 15% of the titles incorporated multimedia content, this proportion is rising steadily. In the near future, multimedia titles in the new related optical formats CD-ROM XA, CD-I and DVI will also begin to proliferate.

STANDARDS

Standards were a serious problem until the recent adoption of the ISO 9660 standard, as well as the appearance of the Microsoft CD-ROM Extensions (MSCDEX) to DOS. It is now possible for industry observers such as Nancy Herther (1991) to view standards in a different light: "Standards continue to be the factor buttressing and fostering growth for CD-ROM." Nevertheless, other observers such as John Dvorak (1991) points out that hardware compatibility problems definitely remain so far as 486 machines are concerned.

Software proliferation has raised another type of standards problem, the lack of consistent search interfaces. More than 100 search software packages are now employed on CD-ROM products, although a core of 15-20 of these account for perhaps 60% of all available products. Although it is not clear that a standard interface is actually desirable (at least not until we find a standard user and a standard database) this situation certainly has raised problems of several types:

After the tenth disk I wondered if the vendors are missing the point. Gads. Each one has a different file format and a different search engine. Each vendor wants you to dedicate a subdirectory and a bundle of megs to a given disk, and each disk requires a screwball path statement, a driver, or some memory-chewing crapola to operate properly. If you routinely used more than a few CD-ROM disks, your system would be chewed up by ancillary files and memory hogging TSRs. (Dvorak 1991)

These problems pose a particular challenge to multiple database workstations, public-access CD-ROM services and end-user instruction. The networking of CD-ROM databases also carries its own problems; however, when networking is practical, it can suffice to greatly simplify user access, addressing many of Dvorak's expressed concerns.

MULTIMEDIA FORMATS

Meridian Data, California based developers of CD-ROM publishing systems, with sales of US$7.2 million in 1989, expects that about a quarter of their clients will begin using their new CD-ROM XA development system, introduced in 1990. (Meridian 1990) The multimedia industry is expected to develop similarly to CD-ROM: "two years following the sale of production units to developers, the end user market should emerge." (Meridian 1990)

Multimedia applications require more powerful (and expensive) hardware than text-based CD-ROM applications. The cost of a multimedia workstation with a 386 microprocessor, 40Mb hard disk, optical drive, VGA card and audio capability, estimated to be about US$3,000 in 1992. (CD-I 1990)

CD-ROM IN LIBRARIES & INFORMATION CENTRES

It is clear, particularly as earlier barriers to implementation have disappeared or decreased in importance, that libraries have implemented CD-ROMs extensively and will do so more extensively yet. Libraries were one of the earliest markets for CD-ROM products and continue to be one of the largest, as Nelson (1991) observes:

> It is no wonder that libraries lead the marketplace in CD-ROM product acceptance. Once they understood the potential of this precocious six-year-old technology, librarians resolutely tackled the several impediments - especially its high price tag - to full implementation of CD-ROM into library processing and public services units. In doing so, they have rightly been accorded recognition as technological innovators.

It is also clear that these systems have received an enthusiastic acceptance by library patrons and end-users, and that the technology is not likely to be displaced in the near future by some alternative new medium. What remains unclear is whether the libraries and information centres should take the initiative to establish inhouse expertise and facilities to assist with "do-it-yourself" CD-ROM publishing. The potential for these applications are universal and just beginning to emerge.

The face of academic computing is changing; it is now very much data driven. The most basic computer-oriented tasks that teachers and researchers must now involve themselves with is the collection, analysis, annotation and organization of information. Richard L. Nolan suggests that we are in a economic transition, one which is taking us from being an industrial economy to an information/service economy. Nolan (1990) argues that higher education must add information technology to its strategic equations for the 1990's. Wether you agree with Nolan that we are in the midst of a new industrial revolution, fuelled by information technology, there still remains an overwhelming awareness of the importance that must be given to information literacy for our future economic success. The American Library Association's 1990 G. K. Hall Award for Library Literature, was awarded to Patricia Senn Breivik and E. Gordon Gee for their book, "Information Literacy: Revolution in the Library." The concept of information literacy involves examination of how information seeking an evaluation skills intersect with the need to understand and use information technology including software; systems design and hardware; access to information beyond that typically stored or accessed by libraries; understanding the complex interaction among stored or remotely accessed information; and finally one's own production of new information.

Libraries and information centres have the opportunity to be the facilitators of inhouse CD-ROM publishing technology; however, useful and innovative applications must originate from all members of the academic community. For example, the following is a sample of CD-ROM applications that might originate from academic settings:

Archaeological artifacts databases comprised of video images and full-text annotations. Rare or significant local collections protected in backroom museums could become transportable and accessible for research and instructional purposes around the world.

Documentation, research and instruction of native heritage. Databases comprised of scanned images, poems, songs, legends, symbolic writing, full-text analysis and audio tracks.

Geographic information databases which cater to municipal and rural communities where Universities and College reside. Virtually all information which has an underlying spatial aspect to it is a candidate. For example, Master's level projects which create geographic image/text databases related to local urban planning issues, commerce, recreation, tourism, health, pollution, housing, population, etc. may find CD-ROM an ideal medium.

Full-text databases comprised of large transcription projects involving international collaboration of scholars and editors. For example, the international scholars working on the transcription of the complete works of Jeremy Bentham (1748-1832).

Preservation and improved access to local history through the construction of specific photo image and full-text news databases. Students in history and journalism may find their compilation and presentation of full-text and image based research to be more effective, accessible and better suited for secondary analysis.

As the cost of inhouse CD-ROM publishing continues to fall and the tasks associated with inhouse publishing become even more pedestrian, we will see a proliferation of CD-ROM information, research and instructional products from academia.

REFERENCES

Campbell, B. "The CD-ROM industry in Canada: trends, companies and products," Optical Information Systems 10(2): 99-103; 1990.

"CD-I, DVI vie for spotlight at Microsoft CD-ROM conference," IDP Report 11(March 9): 1-3; 1990.

"CD-ROM revenues reach $571 million," IDP Report 11(Mar 9): 3-4; 1990.

Cuadra, C.A. "Portable databases: bridging the gap between online and local databases," Information Today 8(1): 15-16; 1991.

Dvorak, J.C. "CD-ROM: still a bust," PC Magazine 10(1): 81; 1991.

Getz, M. "Economics: storage technologies," The Bottom Line 4(1): 25-31; 1990.

Helgerson, L. "When it comes to CD-ROM, it's still a DOS world, but it's changing," CD-ROM Enduser 2(4): 70; 1990.

Herther, N.K. "CD-ROM today: unanswered questions and unending possibilities," CD-ROM Professional 4(2): 10; 1991.

Herther, N.K. "CD-ROM year six: change, growth and stability," Laserdisk Professional 3(2): 5-7; 1990.

Lawrence, A. "Optical discs finding market: still some hurdles to overcome," Financial Post (June 9): 38; 1990.

"Meridian Data expects one-fourth of customers to upgrade to multimedia," IDP Report 11(May 5): 9-10; 1990.

Nelson, N.M. "CD-ROM growth: unleashing the potential," Library Journal 116(2): 51-53; 1991.

Nicholls, P.T. "A survey of commercially available CD-ROM database titles," CD-ROM Professional 4(2): 1991.

Seymour, J. "Forecast '91: will the CD-ROM market finally explode?" PC Week 7(Dec 24): 47; 1990.

Nolan, R. L. "Too Many Executives Today Just Don't Get It!" CAUSE/EFFECT, Volume 13 Number 4: 5-11; 1990.

[1] Presented at the IASSIST 91 Conference held in Edmonton, Alberta, Canada. May 14 - 17, 1991.

Hands on the Census: Microdata from the 1991 Census of Population in Britain

by Catherine Marsh [1]
Department of Sociology and
Department of Econometrics
University of Manchester

The information dilemma

A basic dilemma confronts the producers of data in the public sector. On the one hand, they face users demanding that information collected at public expense be made available to the research community in ever increasing detail. On the other hand they face those concerned with privacy and confidentiality worrying about the possible release of identifiable information about individuals. In short, they have to reconcile competing moral claims: they are caught in the middle between one group of citizens asserting a right to information and another group asserting a right to privacy.

The audience at this IASSIST meeting, as professionals in the business of information acquisition and dissemination, will have rehearsed the arguments many times about the rights to information. It is important to remember that the arguments for a right to privacy and confidentiality are also strong, but have changed their character somewhat during the last decade, and have had a profound influence on the preparedness of state authorities and census agencies to release microdata to the research community.

The state is no longer the only, and perhaps no longer even the major, collector of systematic personal data on individuals. The expansion of information technology has led to a new private industry of information collection and management. Many companies take as their base the publicly available state collected information such as electoral registration lists. Some then may link to this other published records e.g. on bankruptcies and criminal records. Others specialise in the collation of information from pull-out questionnaires in magazines and so-called guarantee registration cards, others bring together personal financial information from major credit card companies and chain stores. Sometimes the aim of this data collection is to assess the credit-worthiness of individuals. Sometimes it is to facilitate modern direct marketing techniques, targeting prime areas for particular mail-shots. Combined with advances in telephone technology, these techniques have become big business. In Britain, several companies have datafiles containing information on over 90% of households. Most of this information is referenced by a combination of name, address and post-code.

There is widespread public concern about the activities of these commercial companies. The British public is very hostile, for example, to the sale of election registration lists to outsiders (Campbell 1987). In the lead up to the 1991 census, there were media programmes and newspaper articles which expressed worry about the potential of census information about small areas to be linked to other private information databases both of these private companies. When the recent Census (Confidentiality) Bill was being debated in Parliament, certain members suggested amendments which would have made the linkage of census information to fine-grain post-coded information illegal (Computer Weekly 31 Jan 1991). The attempts did not succeed, although plans to release statistics for post-coded areas in England and Wales were modified.

Both the political right and left have turned to the state to protect citizen liberties and rights to privacy. Throughout the 1980s, most European countries, unlike the US and Canada, have enacted legislation to give citizens rights with respect to databanks of information which may be held about them. The Data Protection Act of 1984 gives individuals in the UK rights to find out what computerised personal information any organisation might hold about them, to challenge it if wrong, and claim compensation if they suffer as a result. Owners of machine-readable lists are obliged to register any file of identifiable personal information they hold, and to say for what purposes it is held; the register is open for public inspection.

The Data Protection Registrar attempts to police the activities of the information industry; he has, for example, recently attempted to curb the unrestricted use of address-based information for credit-redlining. However, in some ways, the individualistic focus of the existing data protection legislation weakens it as a tool for those seeking to ensure that information released could not be linked to private databanks. Rights to know are restricted to rights granted to individuals to find out what is held about *themselves*; there is no right for someone wanting to establish how much census information it is safe to release to obtain answers to such critical questions as how many individuals are covered in the databank and which census variables are held which

would be available for matching purposes. The format of the Data Protection register does not enable it to be used as a source for answering such questions. Nonetheless, in the Office of the Data Protection Registrar there exists a team of individuals whose job it is to know precisely who has what computerised information about whom, and to police the workings of the Act; in practice they know in broad outline the information gathering activities of all the major data collection companies.

As well as worry that information may be passed on to the commercial sector, the other major public concern about census and survey data relates the use to which the government itself might put the information. Worries have been expressed in Britain at the time of previous censuses about the passing of identifiable census information to immigration officials or to social security officers. A similar set of concerns were expressed in 1991 over the possibility that census information might be passed to Community Charge officers, responsible for compiling lists of the adult population in order to collect a new and very unpopular flat-rate local tax based on a head count (the "poll tax"). Some compaigners against the tax (e.g. in Tottenham in London) explicitly called for a boycott of the census on these grounds. (It is important to emphasise that the census authorities would never in fact pass on census data to any outside the census office, even to other government departments.)

Taking a comparative perspective, it does appear to be the case that the British public is more sensitive to confidentiality issues and less willing to trust the census authorities than in other countries. The British public is more concerned about privacy than other Anglophone countries; Goyder and Leiper (1985), for example, did a content analysis of letters to the press about the censuses in 1980 and 1981, and found much more concern over issues of privacy and confidentiality in Britain than in US or in Canada.

Given the concern about confidentiality, it is worrying to learn that disbelief in the absolute confidentiality guarantees given by census authorities is widespread. This is illustrated by a recent Gallup survey undertaken in both GB and USA. It documents that the level of trust in the census authorities is low in both countries, but lower in Britain: the survey shows that the British public is much less likely to believe the confidentiality pledges given by the census offices than the American public, as Table I shows.

In short, the general public has a range of worries that census information, gathered ostensibly for assisting government in planning purposes of various kinds, will be circulated to others for purposes which were not declared at the time when the information was collected. In lights of these worries, it is not surprising that census offices do not just hand out microdata on request. It is also not surprising that the census offices that made the decision to release microdata earlier on were more liberal about what they were prepared to release than those trying to make the same decisions more recently.

The resolution of the dilemma with respect to census data

Faced with the dilemma between rights to information on the one hand and rights to privacy on the other, the census bureaus in different countries have made different responses. In general, the English-speaking countries have tended to give primacy to rights to information, and have made census microdata available in various anonymised and sanitised forms, whereas European

Table 1: Perception of census confidentiality in Britain and US

Question: "How confident are you that the Census Office will not release an individual's census information to other government agencies: are you ...

	GB	USA
	%	%
very confident	17	23
somewhat confident	27	44
or not at all confident?"	42	28
(don't know)	14	5

Fieldwork dates: March 1991 for GB; March 1990 for US;
Source: Gallup Political and Economic Index, No. 368, April 1991 (GB) and Gallup Report for USA
Reproduced with the kind permission of UK Gallup.

countries have been much more exercised about rights to privacy and have generally not gone down this public road. Britain, standing as it does with a foot in either camp, has taken a long time to decide which way to go.

The United States of America was the first country to release such public use files in the 1960s. From the time when the population census was first computerised in the US, discussion was initiated about releasing forms of microdata to the research community. Partly because the administrative culture was open to research dissemination, and partly because of the existence of energetic individuals pushing from within the census bureau, "public use files", as they are termed in the US, were released retrospectively for the 1961 census and have become a routine part of census output. They were followed by Canada in the 1970s and Australia in the 1980s. However, Canada and Australia never released as much information, either in terms of sample size, detail of file structure or fine-grain detail of coding schemes as the US public use files (see Marsh *et al.* 1991a for more details). The Australian census microdata in particular contains only the state as a geographic identifier.

Despite similar requests for microdata, many European countries have declined to release samples into the public domain unfettered by limitations placed on uses or users. Academic researchers in Denmark and Sweden may only receive microdata for specified and delimited purposes. Some countries allow local state authorities access to census data but deny similar access to academics: local government researchers at all levels in Italy can have access to microdata, for example, as can regional governments in Spain and some government departments in Luxembourg. In Germany, anonymised census records may only be released to the communities[2]. And some countries only release a small subset of the available information in the form of microdata; in France, for example, only a restricted subset of census variables is released publicly as microdata in sampling fractions varying from 0.1% to 25%. (Redfern 1987 briefly outlines the rules in each European country.)

I am delighted to tell you that for the 1991 census in Britain, agreement has been reached for samples of anonymised records to be released. At present the plans only include England, Wales and Scotland, but representations are also being made to the Census Office in Northern Ireland to grant a similar request with respect to the 1991 Northern Irish census.

Requests from British academics for census microdata go back at least fifteen years. Spurred by interest in the release of public use samples in North America, a commitee of interested academics was convened in the mid-1970s to discuss the possibility of obtaining similiar microdata in Britain for the 1971 census. One problem which emerged early on was the geographers' demand for microdata was for very large samples (10% or more of census records) with very fine grain geography. During the 1970s, it seems that the needs of geographers dominated the requests, and the gulf between what academics seemed to require and what the census offices felt able to release while retaining confidence in the confidentiality of the records was wide.

Furthermore, the early discussions about microdata never progressed very far since the legality of releasing microdata under the terms of the 1920 Census Act was never resolved. There were those inside OPCS who argued that release of microdata was permissible under the terms of the Act (Redfern 1976), but the legality of this move was contested by others at the time.

More concerted efforts were made to obtain microdata from the 1981 census. The White Paper outlining plans for the 1981 census made it clear that the census authorities were prepared to consider reasonable requests for samples of anonymised records. Support for release of microdata was also available from other sources, some of them somewhat unexpected: from the British Computer Society team reviewing security provisions for the 1981 census (HMG 1981a), and from these advocating cuts in the Government Statistical Service who argued that microdata could provide a cheap and flexible substitute for tables (HMG 1981b). There was also interest was shown by census office research staff (Denham 1986). And requests from academics (e.g. Norris 1983) for microdata persisted.

A further committee was therefore convened under the auspices of the Environment and Planning Committee of the Economic and Social Research Council between 1984 and 1985, to try to co-ordinate a request to be put to the census offices to obtain retrospective samples from the 1981 census. Despite receiving evidence from several academics about the value of such data, the commitee never reached the stage of putting a formal proposal to the census offices. There appear to have been several reasons for this. First, the demands of the geographers who wanted fine grain areal information to the detriment of detail elsewhere and the demands of social and policy researchers who wanted full household information, if necessary sacrificing geographical detail were never reconciled. Second, it proved very hard to get agreement from others in the commercial and public sectors to form a purchasing consortium to buy the proposed data. Third, in 1985 it seemed likely that there might be a major 10% household survey undertaken in 1986 which might meet needs more effectively; (in fact this survey was never undertaken). As time furthermore, as time went by, the value of data relating to 1981

seemed to decline[3]. The request for data from 1981 therefore lapsed.

However, user demand did not (Marsh *et al.* 1988). The Economic and Social Research Council therefore renewed its efforts in good time for the 1991 census, and set up a working party to negotiate with the census offices in Great Britain and to present a formal request. This working party undertook some systematic work on quantifying the risks of disclosure from releasing census data, and concluded that the risks were minimal. On the basis of that work, it proposed that the census offices release two different files of microdata, one to meet the needs of those who wanted the maximum geographical detail and one to service those whose prime interests were in household structure. The request was presented in a lengthy report in 1989 (published in Marsh *et al.* 1991) which was favourably received by the census offices. The committee also secured the agreement of the ESRC to shoulder the total costs of the purchase if necessary.

The census offices sought advice from their solicitors about whether microdata could legally be released under the terms of the 1920 Census Act. They were advised that anonymised microdata came under the general heading of a 'statistical abstract', and could be released without changing the law. The Office of the Data Protection Registrar and Liberty (the National Council for Civil Liberties as was) were consulted, and neither had any major objections. The proposal to release microdata was therefore mentioned in the White Paper outlining plans for the 1991 census (Cm 430, 1988), subject to the overriding need to preserve census confidentiality. It was also commented upon by the British Computer Society team who reviewed security arrangements for the 1991 census (Her Majesty's Government 1991); since they had supported the idea for the 1981 census, they gave the idea their blessing. In July 1990, the agreement in principle of the Registrars General to the ESRC request was announced in a written Parliamentary answer.

Background to the British Census

In the United Kingdom, censuses of population are the responsibility of three separate offices: the Office of Population Censuses and Surveys under the supervision of the Registrar General for England and Wales, the General Register Office for Scotland under the supervision of the Registrar General for Scotland, and the Census Office in the Northern Irish Department of Health and Social Services under the supervision of the Registrar General for Northern Ireland. The Office of Population Censuses and Surveys (OPCS) plays a coordinating role in the work of the three offices.

The legal framework for censuses in Great Britain is the 1920 Census Act as amended by the Census (Confidentiality) Act 1991. The 1920 Act makes provision for population censuses to be taken at no greater frequency than every five years. It is enabling legislation which requires there to be a new census order outlining arrangements and plans for content each time. The Registrar-General is given the responsibility for organising the census, the power to take on the necessary staff, and the right to present the final accounts to Parliament.

Under the terms of the 1920 Census Act, filling a census return is compulsory for householders. Many in census offices feel that the compulsory nature of the census puts it in a different moral category when it comes to the release of microdata. The logic of this position is not entirely clear, however, since similar confidentiality guarantees are given to those who take part in voluntary government surveys[4].

The Registrar General is given the duty to ensure that summary reports of census data re prepared.Furthermore,

> "The Registrar-General may, if he so thinks fit, at the request and cost of any local authority or person, cause abstracts to be prepared containing any such statistical information, being information which is not contained in the reports made by him under this section and which, in his opinion it is reasonable for that authority or person to require, as can be derived from the census returns." [Section 4(2)].

The interpretation of this section of the Act was critical for the release of microdata; it turned on whether microdata could be deemed a "statistical abstract".

Both by international standards and by comparison with previous British censuses, the censuses of 1981 and 1991 were fairly slim. In 1991 there were 8 questions about housing and 19 questions about individuals. There has never been an income question nor, since 1851, any question on religion in Great Britain, although there is such a (voluntary) question in Northern Ireland. Enumeration is done on the basis of presence on census night; information is also obtained about the usual place of residence of visitors and about absent usual residents (with a voluntary return if the whole household is absent).

There are no long and short forms on the British census. Sampling is only undertaken at the processing stage. The answers to those questions which are laborious to code (such as occupation, industry and qualifications) are only fully coded for a 10% sample.

When it comes to census output, the principle of dissemi-

nation is best expressed in the motto of the London Statistical Society, first expressed over 150 years ago: *aliis exterendum*[5]. At the time it was first enunciated, it was an unattainable ideal, as the technology for taking censuses and surveys had not progressed very far, and the information was usually published in the form of verbal commentaries, often of a very opinionated form (Cullen 1975). However, the trend in census output has continuously been towards making more and more of the detailed information available, aiming for the user to be free to rework it and interpret it in any way that he or she chooses. First this was achieved by providing more tables. There will be 20 volumes of special topic statistics from the 1991 census and monitors for each county and parliamentary constituency. More recently, increased volume of information have been supplied in machine-readable form; the census offices estimate that 98% of census output nowadays is in electronic form. In particular, the small area statistics are particularly detailed; information is provided down to aggregates of around 200 households in England and Wales (70 households in Scotland) and around 9,000 pieces of information are available at this level. There are also special workplace and migration statistics released in machine-readable format for small areas.

Thus, up to 1991, the output from the British census in forms other than microdata was exceptionally detailed, especially in the amount of machine-readable data made available for small areas. This may also be part of the explanation why the demand for microdata in the past never became overwhelming.

The information to be released

At the time of writing, agreement in principle has been given to a plan to release two different samples of data from the 1991 census. Their broad structure has been agreed, but negotiations are still continuing about the precise details. Any of the details mentioned here could therefore change before release of the data.

The first file proposed is a two percent sample of individuals with full housing information and some limited information about household structure attached. The census offices have been guarded about releasing too much information about other household members for fear of effectively releasing what amounts to a hierarchical file under a different guise. This file will show a geographical scheme identifying large local authority districts or groupings of smaller ones.

The second file proposed is a one percent hierarchical file of households, containing housing information and full information about all the individual household members. For confidentiality reasons this file will classify data only to the 10 Standard Regions of Great Britain.

Both files are to be drawn from the 10% of records which are fully coded. The sample will be drawn systematically from this 10% file which is ordered by county, then by enumeration district and street. The two samples will be drawn without replacement so that there will be no overlapping subsamples.

Measures to protect confidentiality of the information

Great effort is put into ensuring that the data is safe from identification and disclosure. Five different devices are being used.

(i) Sampling of records

Sampling itself is one of the most effective ways of reducing the risk of disclosure, provided that users cannot identify which individuals are selected. The sampling fractions are small (.01 and .02), and, because the geographical identifiers are different on the two files, it will not be possible to combine them.

(ii) Supression of variables

Names and addresses are not entered onto the census computer, and therefore obviously not even available for suppression. Precise data of birth is to be suppressed; age will only be available in yearly bands, and will be top-coded (see below). There will also be no information on the imputed missing household[6], since these are excluded from the 10% coding operation.

(iii) Limiting geographical detail

There were competing claims about the basis of SAR geography. The main choices were:

- local government administrative geography
- health service administrative geography
- political boundaries
- local labour market areas
- postcode sector defined boundaries
- grid square geography

It was not possible to have more than one system, as the small overlaps between many of these different schemes would have led to unacceptably small areas being identified. On general utilitarian grounds of maximum benefit to maximum numbers, local government administrative geography was chosen.

The geographical scheme in the individual level file will identify only those local districts with populations of

120,000 or more; the rest will be grouped into areas of at least 120,000 population. However, all bar one of the metropolitan local authority districts will be identified. On the hierarchical file, the geography will be even coarser. Only standard regions will be identified, with London subdivided into inner and outer areas; this amounts to twelve areas in all, the smallest of which East Anglia with an estimated population of 1.9 million.

Other geographical information is also collected on the census - the usual address of visitors to the household, workplace address, students' term-time addresses and address one year ago. These are to be very heavily restricted; at the time of writing, the proposal is to identify standard region, a same district/different district identifier and perhaps a distance measure as well.

The order of records in the microdata file are to be scrambled to ensure that locality information cannot be obtained from the proximity of one records to another.

(iv) Grouping of categories

A rule is being used to restrict the fineness of the coding scheme for each variable:

> Each category of each variable to be identified on the SAR must have an expected sample count of at least 1 in the smallest geographical area permitted on either file.

Expectations are formed by scaling down the distribution of that variable for Great Britain as a whole. To illustrate, in the individual file, the smallest geographical area identified will have a population of 120,000. The sampling fraction will be 1/50. Thus the cut-off on this file is 50/120,000 times the population of Great Britain of 56 million, yielding 23,3000; this is then rounded up to 25,000. Accordingly, any category of any variable which in Great Britain is expected to have less than 25,000 people at the census will be grouped in with another category in the SAR. In the household file, the smallest region identified is East Anglia, with a population of around 1.9 million people, and the sampling fraction will be 1/100, so the cut-off being used for this file is 2,700 people.

Thus only the univariate distributions are used to identify categories at risk. Some have been worried that it is unusual *combinations* of categories that cause problems: female barristers, small householders in accommodation with a lot of rooms, and so on. However, some recent work suggests that the strategy of worrying only about those categories that are small in the univariate distribution will predict people who have unique combinations of variables extremely efficiently (Marsh *et al.* 1991b).

The variables most affected by the grouping rule are occupation (where the 350 categories in the full coding scheme will probably be reduced to around 250), industry (270 categories reduced to around 200), detailed educational qualifications (103 reduced to 53) and country of birth (reduced to 47 groups). The variables to be top-coded are age (to be grouped into two year bands between 91 and 94 and top-coded 95 and above), hours worked last week (top-coded after 70), and number of rooms in the accommodation (to be top-coded after 14).

Top-coding does not solve the problems of households with very large numbers of individuals; the number of people in a household affects the very structure of a hierarchical file rather than the categories of one variable in it. Several suggestions are currently being explored to solve this problem, including removing all geographic identifiers from households with more than 12 members.

(v) Perburbation

The small area statistics from the British census have always been subject to a degree of random perturbation ('Barnardisation') whereby a random +1, 0 or -1 is added all cells. The comprehensive application of this technique has always been unpopular with users (Marsh *et al.* 1988) and the cumulative effects of the small errors introduced can pose quite serious analytic problems (Senior and Cole 1991). Wholesale Barnardisation is of course not possible with microdata, and the prospect of deliberately adding more noise to many different variables on top of the natural levels of error already existing in the data was resisted by ESRC team negotiating the release of the data. Instead, the census offices have suggested that a small number of individuals in each area be switched with others in a nearby area (Griffin *et al.* 1989), a technique which amounts to adding a small degree of random noise to the geographical identifier, but preserving the rest of the household data intact.

In the USA and Canada, the census offices appoint some sort of panel to oversee the arrangements for protecting confidentiality (Gates 1988). Such a Microdata Review Panel was considered for GB, but, on the recommendation of the Royal Statistical Society, the census offices have appointed one technical adviser instead, to oversee the specification of the SAR files and the general arrangements made for confidentiality. This advisor, a senior academic statistician, will make independent recommendations to the Government Minister responsible for the census offices.

Contractual arrangements

The funders of the project are the Computer Board[7], <footnote text> who are paying for the data, and the Economic and Social Research Council, who are putting up the money for a research and distribution centre to be

established to house the data.

The contract to buy the data from the census offices is regulated in part by legislation; under the terms of the 1920 Census Act, the census offices are obliged to recoup the cost of producing extra tables (which includes the SARs), but are prohibited from making profits on the data which they supply. The ESRC team negotiating the release of SARs attempted to find co-purchasers to enter into a consortium to share the costs, but none came forward. Thus now, the academic purchasers are bearing the full developmental costs of the SARs. The census offices have done their costings on the basis of passing on the full marginal costs of producing SARs; at present the cost of the data seems likely to be around £200,000, but this sum will only be finalised when the file specifications have finally been approved.

The contract to must safeguard the interests of the academic purchasers in their product. Although final contractual details are still being sorted out, we hope that, in return for our payment, we will get full exploitation rights of the dataset, which will be sole rights for a limited period. Since the academic purchasers of the data are bearing the full cost of the production of the files, the census offices will not receive any further royalties when value-added products of the census are passed on to third parties.

The contract to purchase from the census offices will specify that all end-users of this data must give various undertakings about respecting census confidentiality. People will be expressly forbidden to try to identify individuals in the SAR, link them with other sources of data or to claim to have done so. Any breach of this agreement would lead to withdrawal of the SARs. While not undoing the harm caused by a breach, but its threat would probably act as a deterrent since academics would subsequently be unable to publish any information based on the SARs. Various methods are under discussion to register users and keep tract of copies. It is possible that heads of departments will be required to be the data holders rather than individual researchers.

The contractual arrangements between the purchasers of the data and the end users with respect to have yet to be discussed in detail. Academic users wanting the data for research purposes will have free access to the data, but means will be sought to regulate the other ways in which the data may be used, to safeguard the interest of the public funders. A graduated scale of charges seem likely, charging commercial users what the market will bear, perhaps with lesser charges to the rest of the public sector and the voluntary sector. University and college faculty who contract their services as consultants outside the academic sector will be charged for their use of the data.

Disseminating the data to users

The Economic and Social Research Council is funding a centre at a university location to house, disseminate and act as a research focus for this dataset. Invitations to house this centre were put out to several institutions, and three teams submitted tenders. The decision about which will be successful is expected in late May 1991.

Four types of usage are envisaged:

(i) On-line access over the academic network

The Joint Academic Network (Janet) connects all universities and many polytechnics and other institutions in Britain. It is funded by the Computer Board, and free at the point of use to all universities. One important method of giving access to the data will be to mount versions of the data at central locations on the network for easy access by any academic users.

The decision about which software to use for the data has not yet been made. SPSSX seems the most popular candidate for the individual file. The hierarchical file of households could also be set up in SPSS, but this might prove cumbersome, especially since there are two different hierarchies within the dataset: individuals can be grouped into either households of families. SAS, SIR and Oracle are other candidates which might be considered. It is likely that eventually the same information may be held in different formats at different locations; while some may view this as inefficient, from the user's point of view there are great gains in terms of familiarity and ease of use.

(ii) Tables service

The census is a benchmark source of social data. It provides denominators for many different researchers' numerators. However, it is not the prime data source for more than a few researchers. It is therefore important for there to be a service to academics which would provide a service of this kind to other academics, although the demand for this may lead to the need for rationing.

(iii) Customised subsets

Another central task that the service centre will undertake is to extract subsets of variables and cases for different users. This will need to have regard to the media most likely to be demanded; a PC/workstation platform is the most likely here, although demand for data on CD Rom and other media will need to be monitored carefully. There will doubtless be demand for teaching datasets for use in schools and colleges.

(iv) Passing on whole dataset

For efficiency reasons, other academic users would be

encouraged to use the service provided over the academic network. But the whole philosophy of providing samples of anonymised records is that users will be free to port them into their own hardware and software environment and not be restricted by earlier decisions. No restrictions will therefore be placed on others wanting entire copies of the dataset.

Public and commercial users do not in general have access to the academic network. They also tend to be familiar with somewhat different software to that used in universities; in Britain, for example, the most popular tabulation software used by market researchers is a product marketed by Quantime called Quanvert. One of the things that the academic purchasers of the data may want to explore is licensing a commercial agency to provide an on-line service for the commercial sector.

The centre housing the microdata will have responsibility for documenting the datasets, and for computing the sampling errors and documenting these. It will also need to establish a user group, and disseminate information about the database to users, both in electronic media and by hard copy newsletters.

Conclusion

The existence of microdata from the 1991 Census opens up a valuable new resource to British social researchers. Our social research community has grown used to making do with small area aggregates, and drawing inferences from these; the existence of microdata should lead to some interesting work on ecological fallacies. Survey and market researchers will be in a position to design much more efficient samples to locate specific subgroups of the population. We will have a source of microdata on a badly neglected part of the population, namely those who do not live in private households. Research into different means of classifying families and households should blossom given the richness of the hierarchical information available.

Perhaps also we may hope that effort will be made to construct internationally comparable files of census microdata relating to the 1990 and 1991 censuses in different countries. One spin-off of presenting this information to an international audience at this IASSIST meeting might be to stimulate discussion in this direction.

References

Campbell, D. (1987) 'The databank dossier', New Statesman, April 24.

Cullen, M. (1975) *The Early Victorian Statistical Societies*, Brighton: Wheatsheaf.

Denham, C.J. (1986) 'Census microdata in Great Britain: the possibilities', Nutzung von anonymisierten Einzekangaben aus Daten der amtlichen Statistik: Bedingungen und Moglichkeiten, Verlag W. Kohlhammer.

Gallup (1991) Gallup Political & Economic Index, Report No. 368, April.

Gates, G.W. (1988) Census bureau microdata: providing useful research data while protecting the anonymity of respondents, Proceedings of the Social Statistics Section of the American Statistical Association Annual Meeting, New Orleans, Louisiana, August 22-25.

Goyder, J., and Leiper, J. McK. (1985) 'The decline in survey response: a social values interpretation', Sociology, vol. 19, 1, pp. 55-71.

Griffin, R.A., Navarro, A., and Florez-Baez, L. (1989) 'Disclosure avoidance for the 1900 census', U.S. Bureau of the Census, paper prepared for presentation at the 1989 Joint Statistical Meetings, Washington, D.C., August.

Her Majesty's Government (1981a) 1981 Census of Population: Confidentiality and Computing, presented to parliament by the Secretary of State for Health and the Secretary of State for Scotland, Cmmd 8201, London: HMSO.

Her Majesty's Government (1981b) 'Initial study of the Office of Population Censuses and Surveys', Annex to the Review of the government statistical services, Cmmd 8236, London: HMSO.

Her Majesty's Government (1991) 1991 Census of Population: Confidentiality and Computing, presented to parliament by the Secretary of State for Health and the Secretary of State for Scotland, February, Cmmd 1447, London: HMSO.

Marsh, C., Arber, S., Wrigley, N., Rhind, D., and Bulmer, M. (1988) "The view of academic social scientists on the 1991 UK Census of Population: a report of the Economic and Social Research Council Working Group", Environment and Planning A, vol. 20:851-889.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., and Walford, N. (1991a) 'The case for samples of anonymised records from the 1991 census', Journal of the Royal Statistical Society (A), vol. 154 (2), 1991.

Marsh, C., Dale, A., and Skinner, C. (1991b) "Safe data or safe settings: disseminating information from the British Census of Population", paper to be presented to

the ISI meetings in Cairo, September.

Norris, P. (1983) 'Microdata from the British Census', in David Rhind ed. A Census User's Handbook, London and New York: Methuen.

Redfern, P. (1976) Releasing statistics as aggregates (tables) or tapes of anonymised individual data (id). OPCS, London: mimeo, available from the author at 17 Fulwith Close, Harrogate, HG2 8HP.

Senior, M., and Cole, K. (1991) '1981 Census costs Department of Health £190,000!', Manchester Computing Centre Newsletter, no. 183, p.2.

[1] Presented at the IASSIST 91 Conference held in Edmonton, Alberta, Canada. May 14 - 17, 1991.

[2] The Federal Republic of Germany has a policy against the general release of microdata collected by the state. It will not allow its Labour Force Survey to be released for secondary analysis, for example. Since the Statistical Office of the European Commission takes the most restrictive policies of its member states as guidelines on release of microdata, this means that Europe-wide release of the Labour Force Survey is precluded.

[3] In fact, the value of census data to academic researchers does not seem to decline as the census data becomes less timely; if usage of small area statistics is a guide, the usage of these increased throughout the 1980s.

[4] Perhaps this is the reason why two other data sources collected compulsorily by law (the New Earnings Survey and Census of Employment) are only released in aggregates, albeit of very small units.

[5] (Trans.) For others to thresh, or, more colloquially, let someone else work out what it all means! It is still the motto of the Royal Statistical Society.

[6] Households with all members absent on census night and which fail to return a voluntary return to the census offices once they get back home.

[7] The body which funds purchases and support of mainframe computing in British universities. It is soon to become a subcommittee of the main Universities Funding Council.

BID - Bringing Integration to Data

by Karsten Boye Rasmussen [1]
Dansk Data, Denmark.

Abstract
The purpose of data archives is not only to store data materials for posterity, but also - and equally important - to fulfil the needs of the present users. The growing demand for studies along with a number of other factors force the data archives to change their strategy for retrieval and dissemination of studies in order to serve their customers in the best - and cheapest - way possible. This calls for effectiveness and new thinking in the retrieval and dissemination procedures.

This paper outlines the state-of-the-art at the Danish Data Archives (DDA) by describing the format of machine-readable records and the present utilization of machine-readable documentation at the archive. A projection of the growing service rate shows the pressing need for a more effective system for direct servicing of the users, the idea being to integrate the study description, the variable documentation and the data and as the distributing medium use CD-ROMs, as the production costs are reasonable. The remaining problem is obtaining or financing the development costs of a suitable retrieval system. These will most probably be rather high, and as probably a lot of archives have the same need for a retrieval system, it could be a good idea for the archives to cover the development costs between them.

Background: Machine-Readable Records
To ensure the right perspective I shall shortly outline the records kept at the DDA[2]. The data are social science quantitative investigations (typically surveys). After the processing at the archive of the raw data and the (typically) paper documentation (questionnaire, coding instructions, reports etc.) each survey consists of three files:

Documentation of the study (study description)
Rectangular (flat) character data file
Documentation of the variables
The files of the finished survey

Study description
I shall not go into details about the study description. Discussions concerning item numbers, subitem numbers and especially the introduction of new item numbers have been the subject for discussion at many earlier IASSIST conferences. The "Standard Study Description Scheme" has gradually been improved[3], but it is still close to the original scheme agreed upon more than ten years ago[4]. The scheme is used at several archives[5]. At the DDA all study descriptions are maintained in both a Danish and an English version.

The actual format and the many items makes the Standard Study Description scheme a rather complex instrument. The design is specifically intended to be a machine-readable record at the study level. The study description format contains coded as well as text information. It is of importance that there is no limitation to the amount of text in the text fields.

```
TITLE    Title of study
CLASS    Ready-made for analysis or just deposited
ACCESS   There may be restrictions
YEAR     The year the survey was carried out
CASES    The number of cases in the survey
NATION   Which country or countries
DONOR    The depositor of the data
INSTITUTE    Place of employment of the depositor
SPONSOR      Finance (Social Science Research Council)
COLLECTOR    Who carried out the field work
ABSTRACT     Description in free text
KEYWORDS     Controlled vocabulary
```

Areas of the study description

The data file

Data stored at the DDA are kept as rectangular character data files (not as system files). One record per case. Interlinked datasets are either divided into separate files (like tables in a relational database) or patted into one (greatly redundant) file. The data for each observation is placed as a record in the data file. And each record holds information of the variables stored in fields (fixed columns). The data per se will not be of any use without the documentation.

```
02091203060203100390030056602011061103902010
02081210101010102002005110801102110310 3100
02063203030102010330062034408011051103302000
02071203060203030330050052301011041103103001
02071201050202010130020051308011011103306000
01044204050204040440022021208011011103101000
01071201050203010190010054308011051103103001
02051299101010040490023031405011061103203001
02072301010102010190110063208011011103302000
02043204050203030430030012208011041103202010
02042312050203121290011022108011021103203000
01034199101010999910030011105011061103201000
02051204020103010430010012408011051102303000
02061204020103020430010053408011011103303000
01073203010102030390990054408011011103103000
```

Character data file

Documentation of the variables

The character data file needs documentation in order to be of any use in its own right. Without the proper documentation the user is not able to distinguish one flat file from another. In order to exploit the information in the data file we need variable documentation. The variable documentation can be thought of as a relational table where each field describes certain characteristics of the variable. This idea is already in production in several databases. In IBM's SQL database[6] several system tables keep track of the information stored in the database. An SQL database may thus be viewed as a complete data archive, where each study is a separate table. The available tables are then stored in the SYSTABLE-table and the documentation of the fields of each table (data) is stored in the SYSCOLUMNS-table.

The project of making a complete description of the fields in the variable documentation lies outside the scope of this paper, so the list below shows only a selection of possible fields. The list shows that the documentation provided by the most widely used social science data packages (SAS and SPSS[7]) does not have the necessary facilities for a complete documentation. Both packages support only the first part of the field list.

```
NAME    The variable name or number
LABE    Short text description
PLACE   Extract the data from these columns
MISSING     Definition of missing data
FORMAT      Output format / categories

QUESTION    The complete questionnaire text

STUDY   Identification of the study
QID     Identification in original questionnaire
FILTER  Reference to filtering
CODINGSpecial coding instructions
PROCESSING  Unofficial comments
CHECK   Checking performed
CHECKLOG    Generated notes on the checking
...
```

Fields for a variable definition

The most important field for the user is the QUESTION field consisting of free text description exactly as it appeared in the questionnaire. This field provides opportunity for performing full text retrieval. Neither SAS nor SPSS are capable of storing and utilizing the complete questionnaire text. The only field for giving a text description of the variable is the variable label. The label in both packages is in practice[8] limited to 40 characters resulting in definitions like:

```
LABEL  INCOME =

'PRS MON AVE INC MAIN-JB -TAX DK 894-904' ;
```

The variable label, 40 characters

40 characters is not much! The variable INCOME actually covers "Personal average monthly payment on main job, after taxation, in Danish Kroner, April 1988-April 1989". This is certainly not a lot of text for a questionnaire text or a created variable, but even then this cryptic label presented above is a common result when the text is restricted to 40 characters.

This is one of the reasons for the DDA to use the data-archaic format provided by OSIRIS[9]. The format was developed by the ICPSR. The virtues of the OSIRIS codebook format is that there are no limitations to the amount of description for each variable. There can be several lines describing the variables and the categories. There can even be references to notes.

The format uses a sort of tag in the first column of each "card" - yes, it is that old - to identify which portion of the variable documentation is described:

T	Label, columns, missing
Q	Question
X	Explanation
J	Unofficial comments
F	Frequency
C,B	Category text
G	Reference to note
S,E	Introduction
M	Note text

The OSIRIS codebook types

Using the format will result in a variable description looking like the following. This is certainly not as straightforward as the syntax used by SAS or SPSS. But normally this lay-out is not written by human beings. A machine-readable format of this complexity and rigidity is best written by machines. In the actual production of machine-readable documentation at the DDA a pre-processor software is used.

```
T0093 EEC VOTE TODAY          018400020   0100000090000009
Q00930093     If there was a referendum on joining the EEC today would
K0093    you vote yes or no?
X0093    In the Danish Election Study, 1975 "Don't know" is
K0093    included in "4. Don't want to answer".
J0093    VARIABLE BASIS
K0093    PRE 1971: -
K0093    POST 1971: -
K0093    DES 1973: 157
K0093    DES 1975: 90
K0093    DES 1977: 211
K0093    DES 1979: 166
K0093    DES 1981: 259
X0093  2     1971-1 1971-2  1973   1975   1977   1979   1981
K0093  ------------------------------------------------------
K0093    1.    -      -     45     37     30     41     41
K0093    2.    -      -     44     38     50     41     40
K0093    3.    -      -      1      3      3      2      2
K0093    4.    -      -      0     13      2      1      1
K0093    5.    -      -      9      -     15     13     16
K0093    9.    -      -      1      9      1      2      1
K0093    11.   100    100    -      -      -      -      -
K0093  ------------------------------------------------------
K0093    WGT N = 1302   1302   533    1600   1602   3192   1500
K0093
C0093    255801. Would vote yes
F0093    UNWEIGHTED:          2558 WEIGHTED:          3231
C0093    280702. Would vote no
F0093    UNWEIGHTED:          2807 WEIGHTED:          3542
C0093     15003. Would return a blank ballot paper
F0093    UNWEIGHTED:           150 WEIGHTED:           189
C0093     28104. Don't want to answer
F0093    UNWEIGHTED:           281 WEIGHTED:           297
C0093     68805. Don't know
F0093    UNWEIGHTED:           688 WEIGHTED:           931
C0093     20809. Not ascertained
F0093    UNWEIGHTED:           208 WEIGHTED:           237
C0093    260411. The variable is not included
F0093    UNWEIGHTED:          2604 WEIGHTED:          2604
```

Example of OSIRIS codebook[10]

The OSIRIS software itself has not been used for many years at the DDA, only the format lives on. But around this format a great many procedures and programs have been developed for checking the data against the documentation, for presenting the documentation, and for further utilizing the documentation.

Utilization of Machine-Readable Documentation

The main reason for the production of machine-readable documentation is the reuse of information. Safely stored documentation can be used in a variety of ways. This chapter describes the current activities at the DDA. The present status implies

some inexpediencies that will be clarified in details in the next chapter.

Printout
Conversion to other formats for analysis
Retrieval
Dissemination

The utilization of machine-readable documentation

Most of these evident virtues of machine-readable documentation goes for the study level (the study description) as well as the variable level (the codebook).

Printout
In compiling a complete documentation a printout of the study description in a readable format (for human beings) is placed as an introduction to the codebook. Printouts of several studies form a catalogue. As the study descriptions are rather voluminous, a catalogue of study descriptions normally includes only the most important items, but the entries are heavily indexed (by persons, subjects, and keywords). However, it is expensive to produce printed catalogues, and it is difficult to keep up with the immediate obsolescence of the catalogue.

To make a printout at the variable level is rather straightforward. The problem is to make the documentation as accessible as possible to the user. Most of the inconvenience of the rigid OSIRIS format is overcome by getting rid of the redundant information in the codebook[12]:

DDA-0658 Danish Election Studies, Continuity File 1971-1981

VAR. 93 EEC VOTE TODAY

start pos. 184, missing data: = 9 or >= 9

If there was a referendum on joining the EEC today would you vote yes or no?

In the Danish Election Study, 1975 "Don't know" is included in "4. Don't want to answer".

	1971-1	1971-2	1973	1975	1977	1979	1981
1.	-	-	45	37	30	41	41
2.	-	-	44	38	50	41	40
3.	-	-	1	3	3	2	2
4.	-	-	0	13	2	1	1
5.	-	-	9	-	15	13	16
9.	-	-	1	9	1	2	1
11.	100	100	-	-	-	-	-
WGT N=	1302	1302	533	1600	1602	3192	1500

unweight MD

marg % %

2558 28 39 01. Would vote yes
2807 30 43 02. Would vote no
150 2 2 03. Would return a blank ballot paper
281 3 4 04. Don't want to answer
688 7 11 05. Don't know
208 2 . 09. Not ascertained
2604 28 . 11. The variable is not included

9296 100 99 Weighted respondents: 11031

Conversion to other packages (SAS and SPSS)
Very few users - if any - make their analyses on delivered datasets using OSIRIS. Instead most users in Denmark use SPSS or SAS. To the user the OSIRIS codebook produces only a printed codebook. The ability of software programs to read system files from other packages has appeared to be quite unstable. With every new release and every new operating system platform this conversion often proves to be the tricky part.

However, with the OSIRIS codebook being a machine-readable variable documentation it is possible to convert the documentation file without affecting the data file. The rigid format of the codebook makes this a relatively straightforward process. Presently the DDA uses a micro computer program called OSI-SPC[13] for doing the conversion. The idea is to leave the data file unaltered. The same data file can be described by both OSIRIS, SAS and SPSS. The user then receives the OSIRIS documentation and the data file, and with the help of the conversion program, the user is able to convert the documentation to the preferred package. And the conversion will produce a SAS-program or SPSS-controlcards, which the user would otherwise have had to write himself.

```
TITLE   'Danish Elections, Continuity file 1971-1981' .
DATA LIST           FILE='L:\U0658\MINI.DAT'
                    FIXED TABLE /
                VAR3 8 VAR93 184-185 .
VARIABLE LABELS VAR3 'SURVEY ID
                VAR93  'EEC VOTE TODAY
VALUE LABELS
        VAR3 1      'Danish Pre-Election Study, 1971'
             2      'Danish Post-Election Study, 1971'
             3      'Danish Election Study, 1973'
             4      'Danish Election Study, 1975'
             5      'Danish Election Study, 1977'
             6      'Danish Election Study, 1979 (no. 20)'
             7      'Danish Election Study, 1979 (no. 21)'
             8      'Danish Election Study, 1981' /
        VAR93 1     'Would vote yes'
              2     'Would vote no'
              3     'Would return a blank ballot paper'
              4     'Don t want to answer'
              5     'Don t know'
              9     'Not ascertained'
             11     'The variable is not included' .
SAVE                FILE='L:\U0658\SPSSMINI' .


SPSS/PC+ generated program from OSI-SPC
```

```
TITLE 'PRSETUP  Two variables from DDA-0658' ;
TITLE2 'Danish Elections, Continuity file 1971-1981' ;
LIBNAME LIBRARY 'L:\U0658' ;
PROC FORMAT LIBRARY=LIBRARY ;
VALUE  VAR3L      1 = 'Danish Pre-Election Study, 1971'
                  2 = 'Danish Post-Election Study, 1971'
                  3 = 'Danish Election Study, 1973'
                  4 = 'Danish Election Study, 1975'
                  5 = 'Danish Election Study, 1977'
                  6 = 'Danish Election Study, 1979 (no. 20)'
                  7 = 'Danish Election Study, 1979 (no. 21)'
                  8 = 'Danish Election Study, 1981' ;

VALUE  V93L       1 = 'Would vote yes'
                  2 = 'Would vote no'
                  3 = 'Would return a blank ballot paper'
                  4 = 'Don''t want to answer'
                  5 = 'Don''t know'
                  9 = 'Not ascertained'
                 11 = 'The variable is not included' ;
       RUN ;
       LIBNAME  U0658  'L:\U0658' ;
       FILENAME DATAIN 'L:\U0658\MINI.DAT' ;
       DATA U0658.MINIDAT ;
       INFILE DATAIN LRECL=185 ;
       ATTRIB     VAR3       LABEL = 'SURVEY ID          '
                             LENGTH = 3   FORMAT = VAR3L. ;
       ATTRIB     VAR93      LABEL = 'EEC VOTE TODAY     '
                             LENGTH = 3   FORMAT = VAR93L. ;
       INPUT    VAR3 8 VAR93 184-185 ;
       RUN ;

SAS setup generated by OSI-SPC
```

The development of OSI-SPC is part of the archive policy. The OSIRIS codebook format is used because of its unlimited capabilities for storing text description. With the help of conversion program(s) the needs of the users of the future can be fulfilled as well. To support new analysis packages we need only to make adjustments to the software (contrary to developing a new conversion program) to stay in pace with the development of new software packages.

Documentation retrieval

I reckon that all archives make use of some kind of retrieval system (possibly many systems) in order to search and identify datasets of interest to the users. At the DDA we have made several systems at the dataset level as well as at the variable level.

The system used for searching study description excerpts (DDAGUIDE) has the most extensive documentation and it is the

only system available to outside users. DDAGUIDE is running on a mainframe host[14], and uses a dialect of the Common Command Language (CCL, promoted by the EEC). Retrieval at the study level is the first step in searching for a suitable dataset:

```
FIND YEAR>1981 CASES>1000 WORD=TV. INDEX=ELECTION

YEAR>1981              Study carried out after 1981

CASES>1000             With more than 1000 cases

WORD=TV.               Words beginning with "tv"

INDEX=ELECTION         "Election" found as a keyword
```

DDAGUIDE retrieval in study descriptions

This request can be regarded as a construction of 4 sets, which are all combined (with logical AND) in a resulting set (possibly empty). Normal Boolean logic (AND, OR, ANDNOT) can be applied for building new sets. The result is a vector of study numbers, and then the codebooks for these studies must be searched to secure that they contain variables that come close to the user's need.

Dissemination of studies
At present studies are distributed to the user by means of many different media: mainframe tapes, diskettes, and electronic network as preferred by the user. Compared with the situation only a few years ago a lot of the analysis is now taking place on micro computers, and consequently many users prefer the data to be delivered on diskettes.

The Pressing Need for more effective Servicing
The growing demand for delivery of studies forces the DDA to make the retrieval and selection as well as the practical dissemination more effective. In this chapter I shall take a look at the present technical obstacles that make the service less effective.

Both the retrieval and the dissemination are carried out at the DDA and require a lot of manpower at the archive. About 25 pct. of the work of the archive is devoted to servicing[15], and because of the growing rate of data deliveries this number is expected to increase. Because no extra funding is available the success of data deliveries will drain the potential for processing new studies to becoming available in fully documented form. It is thus a necessity to make the retrieval and dissemination procedures more effective.

More effective retrieval tools
The present retrieval tools suffer from the following drawbacks:

Dispersed retrieval facilities

DDAGUIDE searches only study descriptions

Updating is irregular and cumbersome

Retrieval only available on one specific mainframe

No thesaurus (actual words searched)

Line mode user interface

Drawbacks of presently used retrieval tools

Having several poorly integrated and sometimes individual retrieval facilities makes it more than a one-man-job to select the appropriate studies. Using DDAGUIDE will narrow the number of studies to be investigated further, but these studies will have to be scrutinized at the variable level. This means that the codebooks must be searched, but at present no system is available for searching the codebooks, therefore the most obvious choice is to ask the people at the archive responsible for the selected studies. Secondly, some studies (series of studies like Gallup polls etc.) have some extra machine-readable indexing of variables, but these indexes are not integrated into the codebooks and are made by a third person. This shows poor integration and a massive use of manpower, which eventually introduces the risk of performing erroneous retrieval.

The solution is first to integrate the index with the codebooks. The retrieval of codebooks must then be integrated with the retrieval of study descriptions. It ought to be possible to select a set of studies at the study level. Furthermore the variables of these studies should be searched within the same system. Superficially there is not much difference whether the unit of retrieval is a study or a variable. But this proves to be wrong, as this example illustrates[16]:

SET1: Housing

SET2: City

1 AND 2: Variables with both "Housing" and "City"

VAR: found within the same single variable

STUDY: found 2 variables within same study

Combination (AND) of retrieved variables

A retrieval system for codebooks has to have a kind of extra parameter to distinguish whether the combinations of variables take place at the variable level (both subjects found within the same variable) or at the study level (two variables covering both subjects found within the same study).

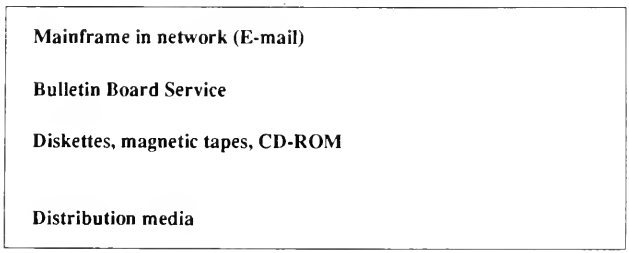
Updating text databases presents major problems. Adding new studies or variables is no problem, the obstacle exists in replacing a text entry. This is caused by the most used algorithm where the actual text is subdivided and scattered throughout the data base. The simplest solution is to build the data base from scratch every time replacement or updating is required. This demands computing power, which in turn is becoming still less expensive. Earlier retrieval data bases were typically placed on large mainframe hosts. The performance of recent microcomputers makes this kind of machine a perfect choice for the integrated retrieval system.

Unfortunately an integrated system for retrieval at both the study and the variable level is not available at the DDA at present. Some of the design - mostly in the form of wishful thinking - is presented in this paper, but the actual development requires funding. We are not so stubborn as to insist on making this development. Presently we have not come across a system fulfilling our demands, but we try to experiment with systems and we are looking forward to the perfect system being presented. Also a lot of archives must be investigating into the same problems, so maybe a more integrated and common effort is required to reach the goal.

Bringing the Data to the User

If we assume that the user has decided on which studies he wants, the distribution of studies to the user can take place in a lot of different ways.

Presently the DDA places the studies on the requested medium (diskettes, mainframe tapes or using electronic mail). This demands active work from the employees at the DDA. But in addition to sending data to the user there exist several other distribution methods that require a more active role of the user.

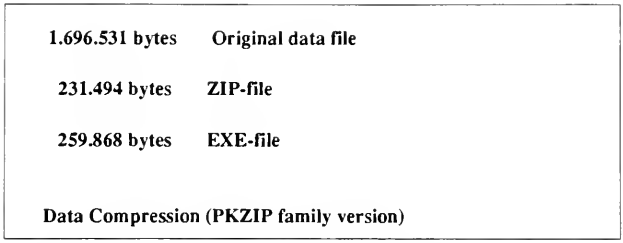


Mainframe in network

If the users are all connected to the same machine the solution is simply to place the archive files on this computer. This is the solution of campus-archives, but it is seldom efficient for archives with national coverage. However, as mainframes are being connected with networks this solution has turned out to be a big success. Viewed from the ICPSR, access through network (CDNet) in 1989 "accounts for almost three quarters of the ICPSR data orders"[17]. The success is totally depending upon the users' access to, knowledge of and familiarity with the networking procedures. Although the DDA is the national data archive for Denmark, in a lot of respects it would be a great mistake to compare the DDA and the ICPSR. In this context the big difference is that the ICPSR is serving trained staff at member institutions. This staff is then handing out the data to the users, which on their side are accustomed to using the campus computer. At the DDA we are serving the user directly. The users are using a lot of different mainframes and are seldom aware of nor interested in the networking techniques.

Bulletin Board Service

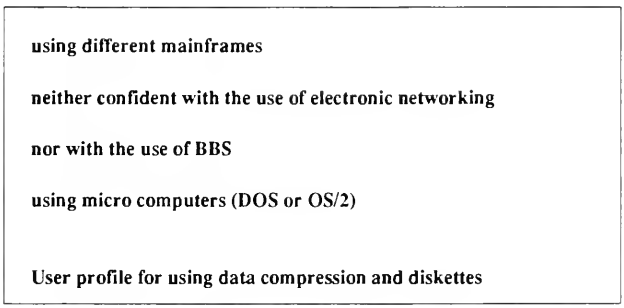
All the users can be expected to be using some kind of micro computer. It would seem rational to make the data archive holdings available on a remote basis for PC-users. The most common form to be utilized is then running a Bulletin Board Service (BBS) with download facilities. However, most of the users can not be expected to possess the necessary technical facilities (especially modem connections). As a side effect, the investigations into BBS have drawn our attention towards data compression techniques. In a recent BYTE article comparing data compression software[18] the virtue of data compression is seen as a saver of space on the hard disk, but for long the greatest potential for data compression has been the dissemination of data via BBS by modem (normal modem without built-in data compression facilities). However, the data compression technique is fully applicable to diskettes. This example shows what is saved by using the PKZIP[19] data compression:



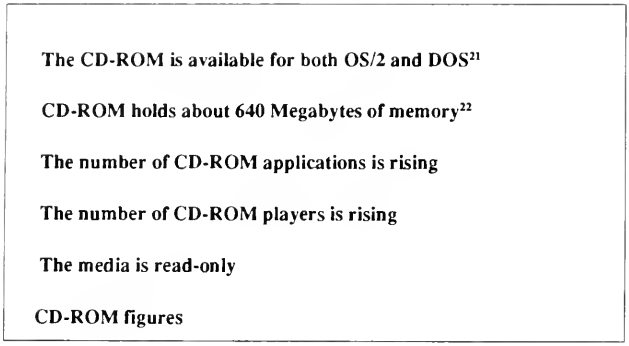
Due to the limited variation of the bytes in the data file the compressed data file gains 86 percent of the space required by the original file. This means that a file that is too big for a single one of the largest diskette formats available (IBM 1.44 Megabytes) will now fit an old floppy disk (360 Kilobytes). The ZIP-file is converted back to the original format by an unpacking program. To make sure that the user will be able to unpack the ZIP-file - even if the user does not possess the UNZIP program - the ZIP-file can be extended to a self-unpacking program (an EXE-file produced by the ZIP2EXE program). This will cost around 30.000 bytes. Furthermore there exists a family version (a program that is capable of running both under DOS as well as OS/2), so the produced EXE-file can be unpacked in any of the two environments.

By using the most common baud rate (2400 baud) it would take approximately 20 minutes to download the EXE-file.

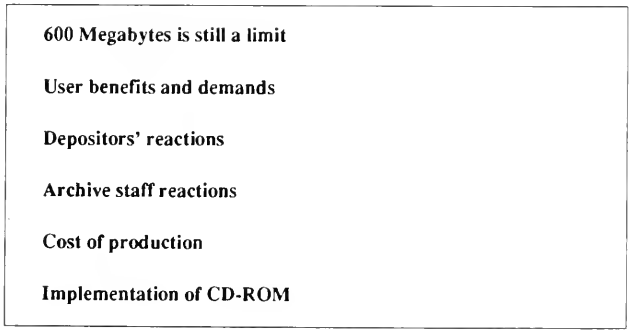
Diskettes with data compression
Presently data compression combined with the mailing of diskettes seems to be the best alternative for the dissemination of a minor collection of studies to users:



The Data Archive on a Disk (CD-ROM[20])
The real potential of a data archive lies in the user's opportunity to perform comparative analyses on more than a single study. This means that most secondary analyses will need several studies, and that a single diskette will prove insufficient. A radical solution would be to put the complete archive on a disk which could be distributed. This has become relevant with the marketing of optical disks. Also WORM-disks exist in a lot of different and incompatible formats. But with the growing number of CD-ROM players the CD-ROM medium seem to be the standardized and perfect solution for bringing out the archive to personal computers.



Normally read-only media are thought of as some kind of second class media, but read-only is the perfect attribute for the stable archive data. In this chapter I shall present some storage considerations as well as some different viewpoints on the implementation of a CD-ROM solution.



Is 600 Megabytes enough?
Lets take a look at the size of data stored at the DDA. In the following table the unit is files archived at the DDA. But each file is in addition weighted with the number of bytes that it occupies. Totally we are talking about 5398 files occupying 2953 Megabytes or 3 Gigabytes. The main archive is placed on mainframe tapes.

Each file belongs to one of the categories: Finished, Process (being processed, i.e. an intermediary file that is stored for extended security) or Original (the file received). Likewise each file also belongs to one of the package-categories (OSIRIS, SPSS or SAS) and finally the file contains either DATA or documentation (DICT/DICB/MISC)[23].

Of these files only a small fraction of the 3 Gigabytes are of interest to the CD-ROM project namely studies finished and stored in an OSIRIS version: At present 586 data files (341 Megabytes) together with the documentation files (491 DICBs (dictionary-codebooks) totalling 94 Megabytes and 546 DICTs totalling 6 Megabytes[24]). At present these approximately 500 finished studies will occupy around 450 Megabytes. A CD-ROM will hold from 540 to 640 Megabytes.

		STATUS						All	
		Process		Finished		Original			
		N	SUM	N	SUM	N	SUM	N	SUM
OSI									
	CDBK	21	4 Mb	.	.	20	4 Mb	41	9 Mb
	DATA	480	412 Mb	586	341Mb	1412	1417 Mb	2478	2171 Mb
	DICB	21	12 Mb	491	94 Mb	125	41Mb	637	148 Mb
	DICT	302	3 Mb	546	6 Mb	227	6 Mb	1075	16 Mb
	DIV	448	158 Mb	40	11 Mb	80	48 Mb	568	218 Mb
	ALL	1272	590 Mb	1663	455 Mb	1864	1519 Mb	4799	2565 Mb
SAS									
	DIV	2	0 Mb	9	0 Mb	1	0 Mb	12	0 Mb
	ALL	2	0 Mb	9	0 Mb	1	0 Mb	12	0 Mb
SPSS									
	CTRL	30	2 Mb	90	9 Mb	104	9 Mb	224	21 Mb
	DATA	13	35 mb	83	133 Mb	146	184 Mb	242	353 Mb
	DIV	7	0 Mb	5	0 Mb	109	11 Mb	121	12 Mb
	ALL	50	38 Mb	178	142 Mb	359	206 Mb	587	387 Mb
ALL									
	CDBK	21	4 Mb	.	.	20	4 Mb	41	5 Mb
	CTRL	30	2 Mb	90	9 Mb	104	9 Mb	224	21 Mb
	DATA	493	447 mb	669	475 Mb	1558	1602 Mb	2720	2525 Mb
	DICB	21	12 Mb	491	94 Mb	125	41Mb	637	148 Mb
	DICT	302	3 Mb	546	6 Mb	227	6 Mb	1075	16 Mb
	DIV	457	158 Mb	54	12 Mb	190	60 Mb	701	231 Mb
	ALL	1324	629 Mb	1850	598 Mb	2224	1725 Mb	5398	2953 Mb

At least 100 Megabytes of the space on a CD-ROM would be left free. And with compression techniques the required space would be less than 100 Megabytes. So a CD-ROM disk could hold approximately 2000 surveys and still have plenty of free space.

Bringing retrievable documentation to the user

For each study the documentation should be available to the user as well. But the user - and the archive staff - need a retrieval system integrating the documentation at the study level as well as at the variable level as mentioned above. The retrieval demand is the reason for setting aside some free space on the CD-ROM for the distribution of retrieval software, indexed files, and other assisting materials.

Integration of study and variable level documentation

Subsetting of variables

Conversion to other packages

Logging of search criteria

Inclusion of search criteria

Human interface

User needs for integrated retrieval

Even though a CD-ROM can hold a lot of information there is no guarantee that the user has similar abundance of space available on his hard disk. Often the user will need only selected variables from a study. In addition to this type of selection, the user should have the option of transferring the background variables. This calls for the possibility to mark off the background variables in the documentation. Apart from hardware limitations there may be some software limitations[25] too. A lot of patience is required when building datasets with a great number of variables on a PC[26].

Retrievals may become quite complicated, so it is a necessity that the user is given the possibility of tracking down the searches and combinations. A log will provide documentation of the actual retrieval for documentation. Similarly a session should be able to start off where a previous session was left.

The line-mode CCL search syntax is developed for dumb terminals. But the increased use of micro computers has accustomed users to more intuitive search systems. Many search systems now employ pull-down menus as well as windowing systems[27]. The graphic user interface standardized in IBM's CUA[28] uses radio buttons, push buttons, check boxes, list boxes etc. as well. At present I have no knowledge of the CUA standard used in an actual implementation of retrieval software, but probably it has already been marketed.

As most users are conservative about learning new computer languages for analysis it should be possible to convert the documentation to the user's preferred package as exemplified earlier in this paper. An integration of a (new) analysis package could be counter productive for the user and would definitely involve some further costs.

No answer to questions like:

"When was the proportion of Social Democrats among men higher than among women?"

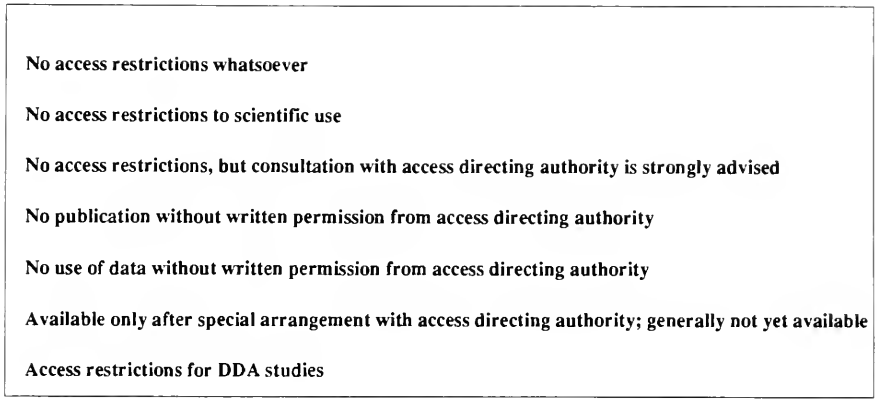
Analysis can be carried out in other processes (OS/2)

With the analysis package preferred by the user

Analysis package separated from retrieval

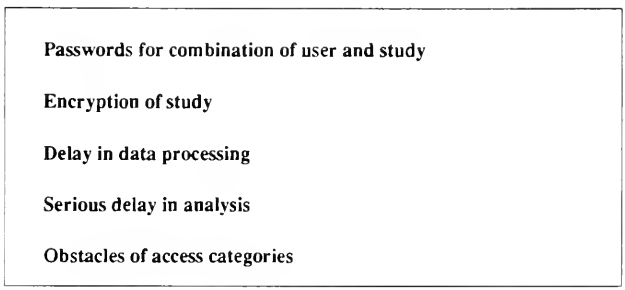
Depositors reaction

The studies stored at the DDA are not all directly available to the user. Although the data are placed at the archive, the depositor still has the formal rights to the material, and the depositor has the possibility of assigning access categories to the dataset. The access categories are assigned to the data only, the documentation is always available without special permits.



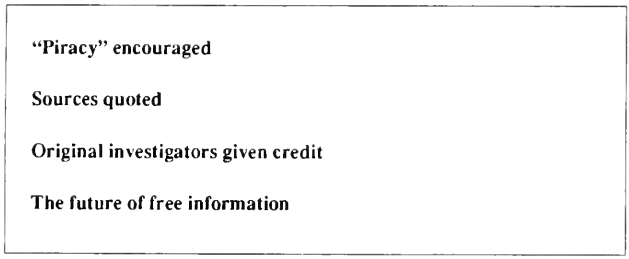
One solution to this problem would be to incorporate only studies without access restrictions (the first three categories) in the BID-project. As this would be a dramatic solution it is not advisable. Another solution would be to work for a transfer of all studies to the category of free access. But this work has already been carried out in so far that most studies after a while are transferred to a less restrictive category. But even though the data are freely available, the depositor (the access directing authority) is presently being informed of whom the datasets are disseminated to. All in all this implies some kind of password, so that it will be possible for the user only to access files that he has been positively assigned to.

The password ought to be distinct for every combination of user and dataset. But user information can not be placed on the CD-ROM, and the password would then have to be distinct only for the dataset. As dataset protection passwords are not a standard feature of the operation systems for micro computers, the passwords would be implemented in the form of a key for the encryption of the data file. A user would then be able to pass-on the key to other users. This shows that the security is not perfect, but then it does not have to be perfect. The same thing is happening now, when a few users are actually distributing the data they have received from the DDA. This is in contradiction with the agreement between the archive and the user and therefore illegal. But unless the analysis system is an integrated part of the retrieval system the spreading of data can not be prevented.



But every restriction introduces drawbacks for the user. First of all the data would have to be both decompressed and decrypted unless a program would be able to do this in a single pass. Secondly - and most seriously - the user would have to contact the archive to receive permission. This would delay the actual analysis by several days for the studies placed in the most restricted categories that requires the interaction of the depositor.

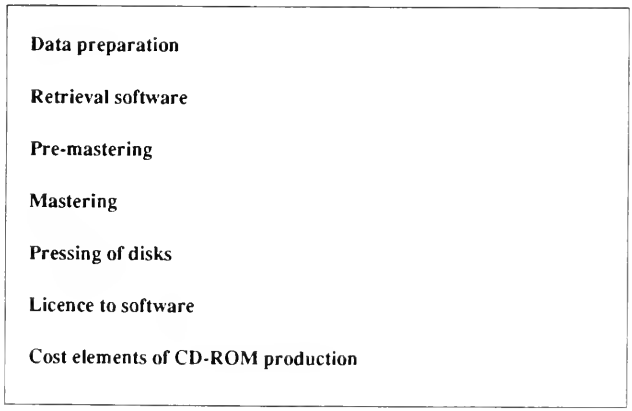
As proven above it is impossible to totally prevent the "pirating" of data. And in my opinion it should be legalized and encouraged. The data and documentation of social science research - which in Denmark is most often funded by the State - should be regarded as public property. Any use of machine-readable material should imply the same consideration as the use of other sources of material. This means that all sources should be quoted, and at the same time the original investigators given credit.



Archive staff reaction
The BID system is intended for the user. But the system should be used at the archive as well. At the archive it would be possible to maintain a more updated version in order to give access to the newest studies.

It must be foreseen that some staff work would be transferred into giving advice on the use of the BID CD-ROM system. But a major part of the present service work at the archive would disappear and leave more resources for bringing up studies to the highest documentation level.

Cost of production
The production costs of CD-ROMs have gone down very fast over the last year. Recent announcements mentions prices as low as 30.000 DKr[29] for the total production of 300 CD-ROMs. The cost would be expected to be even less in the US. But recently a figure of USE 10.000 (70.000 DKr) for the complete process has been mentioned[30].



The difference in pricing is caused by the exclusion or inclusion of some steps in the process. The low costs are obtained if you simply have some files and want them available on a CD-ROM. The receiving company will then do the mastering and pressing of the disks. So this is comparable to the actual printing costs.

If all data and documentation are available the expensive process will be to develop or apply retrieval software. Ready-made retrieval tools for CD-ROM production are available. For companies using computers the software can be acquired, indexes constructed, and the system tested locally. The technical requirements are a medium-sized PC, lots of disk space (maybe optical) and some sort of medium for copying the large quantities of data to the mastering company (tape).

Do not forget that software is under the law of copyright. This means that although you can locally set up a nice system using the software, you are not allowed to put the retrieval software on the CD-ROM. A licence or royalty fee is required in order to distribute the retrieval software to the users.

Conclusion

A lot of different companies and software packages are available to chose among when deciding on the retrieval software. But the "plastic"-CD-software I have seen have all been limited to typical library systems. They could only handle the retrieval from one level of information (Eg. bookcards) and not the hierarchy of studies and variables

Retrieval on two levels (study and variable)

Start of other external processes:
- **conversion to data analysis packages**
- **decrypting the data file**
- **decompressing the data file**
- **subsetting the data file**

Storage of search profiles

The need for open data archive software
***is* The price of integration**

This need for retrieval software that can be further developed is arisen from the idea to integrate the study description, the variable documentation and the data. The conclusion of this paper is that the total integration of these parts makes it even more pressing to make data, documentation and the retrieval software freely available without bureaucratic hindrance.

This paper has been an advertisement for a retrieval instrument open for further development and not the introduction of the product from a concluded development. But the need must be very similar at many data archives, and there should be a potential for covering the development costs between the archives.

Footnotes:

[1] Presented at the "IASSIST 90" Conference held May 30 - June 2 at the Radisson Hotel, Poughkeepsie, New York, U.SA.

[2]. The DDA documentation standards were earlier presented in my paper "Data on Data", in "SUIGI'89", Proceedings of the SAS European Users Group International Conference, SAS GmbH. The BID project has in an earlier version been described in "Beskrivelsens Integration med Data" in DDA-Nyt 48 (in Danish).

[3]. "Standard Study Description Scheme". Latest update 1988, available from the DDA (DOC00364)".

[4]. "Study Description Guide and Scheme" by Per Nielsen, Copenhagen, DDA, 1975.

[5]. The use of the standard study description - and similar vehicles - among European archives described in: "From Localization to Cataloguing of Data Sets". (Workbook of the First CESSDA Expert Seminar). Ed. Astrid Bogh Lauritzen, 1987, Danish Data Archives, Odense, Denmark.

[6]. "IBM OS/2 EE 1.1. Database Manager Programming Guide and Reference", 1988, IBM. (Appendix C). The Database Manager is part of the Extended Edition for OS/2.

[7]. "SAS Language Guide for Personal Computers. Version 6 Edition". 1985, SAS Institute Inc., Cary, NC, USA. "SPSS/PC+ version 2". 1988, SPSS Inc., IL, USA. The Mainframe versions of SAS and SPSS do not differ significantly with respect to documentation facilities.

[8]. "SPSS-X User's Guide" 3rd edition, SPSS Inc., Chicago. SPSS-X supports up to 120 characters, SPSS-PC+ (now version 3) places the limit at 60 characters, but most procedures print out only the first 40 characters.

[9]. "OSIRIS III. Volume 1, System and Program Description" 1973, University of Michigan, USA. (Appendix D describes the OSIRIS dataset).

[10]. This codebook is untypical for studies at the DDA. First of all it is translated into English. The codebooks at the DDA are mostly in Danish. Secondly the study actually consist of 7 studies that have been merged, this accounts for the tabulations showing response- percentages at different points of time. The study is now available through ICPSR (DDA-0658 "Danish Election Studies, Continuity File 1971-1981", ICPSR-8946).

[11]. The latest versions of the catalogue of holdings at the DDA are: "Danish Data Guide 1986" and "Danish Data Guide Update 1988". They are both in English.

[12]. The reason for the atypical layout is explained in note 9.

[13]. OSI-SPC runs under DOS and OS/2 and is available on application to the DDA. The program is capable of subsetting variables from an OSIRIS documentation.

[14]. The DDAGUIDE retrieval database runs under IBM VM/CMS.

[15]. "DDA Annual Report 1988", in DDA-Nyt 49 (in Danish).

[16]. I shall not mention the standard retrieval problems: How to ensure that the employed retrieval terms actually cover the subject searched for; and how - at the same time - to obtain both a high level of precision and a high level of recall. These aspects are covered in "Information Retrieval Experiment", Karen S. Jones (ed.), 1981.

[17]. "ICPSR Annual Report 1988-1989". The ordering of datasets is free but the use of the CDNet database SEARCH is a charged-for service.

[18]. "Saving Space" by Steven J. Vaughan-Nichols, BYTE march 1990.

[19]. PKZIP (PKWARE Inc.), Ver. 1.01 DOS and OS/2 family mode. The latest and much faster version is 1.10 for DOS only.

[20]. The CD-ROM "bible" is: "CD ROM. The New Papyrus", Microsoft Press, 1986. In this collection of early papers Leonard Laub's "What is CD ROM?" is recommended as an introduction to the media.

[21]. IBM has just introduced SCSI-interface in their PS/2 family and also marketed a CD-ROM player supporting this format.

[22]. The CD-ROM maximum capacity is from 540 Megabytes to 640 Megabytes depending on the software used for processing and the software (and hardware) used for reading.

[23]. The study descriptions are left out of this calculation, as their size is relatively unimportant.

[24]. A few of the finished studies do not include a codebook, only the dictionary (variable location and label etc.)

[25]. The SPSS PC+ DATA LIST can not read more than 200 variables nor read ASCII files with a record length of more than 1024 bytes. ("SPSS/PC+ version 2". 1988, SPSS Inc., IL, USA. page C-38). In practice PC SAS (DOS) has limitations to the number of variables too.

[26]. Both SAS and SPSS are just now being released in OS/2 versions that do not have the memory problems of the PC-DOS version. Thus software limitations concerning the number of variables will disappear as the operating systems demand more hardware.

[27]. The ready-to-use software guides from Norton and Microsoft also have the potential for setting up new user defined retrieval systems. WordCruncher from Electronic Text Corporation is especially designed to search great quantities of text information.

[28]. "SAA Common User Access Advanced Interface Design Guide", IBM, 1989 (SC26-4582-0).

[29]. "Compact Data Nyt", April 1990 (Newsletter in Danish).

[30]. "Do-it-Yourself CD-ROMs" by Wayne Rash Jr., BYTE May 1990

The Promise of Multimedia: Data for Every Computer

by Janet Vavra[1]
Inter-university Consortium for Political and Social Research

When the microcomputer arrived on many of our desks in the 1980's, few of us realized just how dramatically this relatively small (often mysterious) piece of equipment would change our lives. Not only did the microcomputer alter the way we perform our daily tasks, it literally changed the way we view and interact with the world. Sending messages to colleagues half a world away was something we did either through Western Union or the postal system; data were shipped on magnetic tape through the mail. One looked for a response to an overnight letter in several days, assuming the marine life did not nibble through the undersea cable in the meantime. Today with the use of email facilities and public data networks, individuals send messages across the country and around the world with the same ease as they once picked up the phone and placed a call across town. Today many users have desktop computers (workstations) that have more computing power than many mainframes had 10-15 years ago, and the machines do not require an entire floor to accommodate them and their peripherals.

The user of a microcomputer, or workstation, can work with an enviable array of hardware and software performing data transfers, database searches, accounting tasks, data analyses, and receiving and sending messages without ever leaving the office. A researcher could conceivably conduct a project from beginning to end from the keyboard (or with a mouse) of a micro.

By searching the library's on-line catalog, relevant publications can be identified thereby eliminating long hours spent in the library by the researcher or a graduate student. Searches are not necessarily limited to local libraries, but rather identify publications available from a variety of sources. On-line databases can be used to locate machine-readable data containing the needed variables. If access to such databases is limited to selected users, a request asking for a given search could be sent via email to the individual authorized to search the databases.

Data could be ordered through electronic mail or electronic ordering systems. In some instances data could be downloaded through public data networks directly to the user's hard disk or to a local data library facility and then through a local area network to the user's machine.

Analysis can be performed with micro-based software and the results shared with project colleagues who may be at other locations, again using email or public data network capabilities. Small files of information can be exchanged through Bitnet, while larger files can be sent through networks such as Internet.

Finally research reports can be prepared with wordprocessors, many of which have graphics capabilities. Frequently the ease with which changes can be made in any document can create a problem of another sort: one cannot resist making another "final" change or one more addition to the document.

One could go on citing similar scenarios in almost every field and activity. The point is that many of the capabilities that were only available to users at central computing facilities, or not at all, are now available on the powerful machines many people have in their offices. Needless to say, this has impacted on organizations, such as the Inter-university Consortium for Political and Social Research (ICPSR), that provide data and related services to many of these users.

This paper will try to identify some of the changes brought about by the advent of the microcomputer. It will primarily look at those changes that impact both on researchers and on those organizations that provide machine-readable data to researchers and instructors. It will seek to identify some of the ways in which these organizations will and may provide their services in the future. The focus will be primarily on the challenges faced by the ICPSR both in continuing to serve users with more traditional computing environments and those at institutions where the activities of the central computing facility have been largely replaced by personal computers.

While the past two to four years have seen a growing interest in what can be termed "alternate media", data continues to be transmitted and exchanged among facilities on magnetic tape. However, the day is rapidly approaching where magnetic tape will not be the medium of choice but rather the medium of last resort. Magnetic tapes continue to be a medium that generally requires a mainframe. This is at a time when many central computing facilities are starting to cut back on services they have traditionally

provided, and many are also cutting back on staff that support these services. The reasoning frequently is that mainframes will serve as giant gatekeepers and servers while the microcomputers will take over the day-to-day computing needs of users.

With the central computing facilities cutting back on individual user services and users finding themselves with impressive computing power on their desks, it is natural that demand will increase for data and other supportive services that are compatible with micros. However, given the variety of configurations one can have at the micro level and the differences in individual preferences, it is no easy task to come up with products that will meet everyone's needs. Additionally archives face the very painful reality of the very high cost of converting all of their holdings from a mainframe-compatible format to a primarily microcomputer-compatible one.

Almost without exception, microcomputers have floppy diskette capabilities. However, not only do the diskettes generally come in two different sizes, they also can each be written in different densities. (Sounds a bit like the old days of seven- and nine-track magnetic tapes.) One can try to identify the format used by the most users and write diskettes routinely with those specifications. While that may be the only thing that makes sense for an organization that supplies data to thousands of users each year, there will remain those users who absolutely cannot use the standard product and must have data at different specifications. It seems to make sense to have a standard product that most users can work with and to deal with those users that cannot handle the standard product on a case-by-case basis.

Unfortunately, while floppy diskettes are an excellent medium for transmitting small collections of data, problems begin to arise when large data collections that contain more than a couple of megabytes of data are involved. One solution is to supply the data in compressed format. Most of the compression software around reduces the size of a file from 70%%-80%% of its original size. Since the capacity of most diskettes ranges from an average of 350 kilobytes for a 5 1/4" low density diskette to 1.4 megabytes for a 3 1/2" high density diskette, it is easy to see that large files, even when compressed, very quickly become impractical for this medium. Supplying one data file on numerous diskettes can create problems for software that may eventually have to manipulate the data. Large compressed files have to be decompressed and space has to be available locally to accommodate the decompressed data.

While the user has to be concerned with the amount of space available on their their microcomputer in order to be able to work with the data arriving on diskette, the data producer is further concerned with the effort that must go into preparing the data for diskette. It may be necessary to reformat the data to make it compatible with the micro environment. For example, PC-based software frequently cannot accommodate large record sizes with ease. Additional problems may arise with large numbers of cases and/or large numbers of variables. Depending on the work that needs to be done, reformating could be an expensive proposition. Accordingly, it may be necessary to identify only certain collections that can be provided on diskette and further to routinely provide these data in only a selected number of diskette formats.

Optical media go a long way toward solving storage problems when it comes to large collections of data. One of the more popular optical media is the CD-ROM. On a disk no larger than a 5 1/4" floppy, a CD-ROM can easily hold over 600 megabytes of raw, ASCII data. While access on a CD-ROM is slower than with a floppy, many producers bundle the data on their CD-ROM with software which helps to reduce the retrieval time. In other instances, users are not concerned about the length of the retrieval time, since time spent on their microcomputer does not result in direct costs the way time spent on a mainframe does. Instead they set up a batch job to run on their micro during the lunch hour or overnight.

Despite the high capacity of a CD-ROM and some of its other attractive features, the CD-ROM is basically not an inexpensive medium. Users usually must purchase a CD-ROM drive. Generally a CD-ROM's performance depends both upon the drive and driver used and on the power of the machine on which the work is being performed. It may even mean purchasing a different micro, if the current micro is not suitable for CD-ROM applications.

From the producer's point of view, a CD-ROM product can be a very expensive undertaking. If the data are to be

bundled with software, the producer must either identify existing software and then seek licensing agreements to use the software, or must write in-house software. Licensing agreements can be costly; the preparation of in-house software may, however, involve an even greater financial commitment. Data will additionally need to be prepared for input into the software or may need to be restructured for the microcomputing environment. Another alternative is to not supply any software with the product and leave it up to the user to identify software to be used with the data. This latter approach is more akin to using the CD-ROM as a data transmittal and storage medium than as a complete data transmittal, storage, and retrieval system.

While data can also be compressed on CD-ROM, the large volume of information that can be stored on CD-ROM usually necessitates special retrieval software for full or partial extracts. It is easy to visualize the problems that could arise if a user had to decompress a 550 megabyte file stored on CD-ROM onto a hard disk, or other local storage media, before being able to manipulate the data.

After the decision has been made regarding the nature of the CD-ROM product, premastering and mastering must be done before copies can be made. Normally premastering and mastering is done by service bureaus although producers can opt to purchase the necessary equipment and software for in-house capabilities. The charges for such capabilities preclude most organizations from deciding to master their own CD-ROM products. Generally the costs for producing a CD-ROM are such that only selected data collections can be considered for the medium.

The transmittal of data over public data networks has a great deal of appeal to both the users and the data producer. By simply identifying the data needed and giving the appropriate set of commands, the user can theoretically transfer any data collection needed in a matter of minutes. This can all be done without any direct intervention by the producer; the producer need only be notified in some electronic manner that the transaction occurred. As network speeds have been increasing from T1 (maximum 1.544 megabits or 200 kilobytes per second) to T3 (maximum 45 megabits or 590 kilobytes per second), this mode of data exchange has created a great deal of interest. But as with all of the alternate media discussed so far, there is good and bad with this option.

It is very attractive from a user standpoint to be able to simply give a few commands on your desktop and have megabytes of data arrive over the lines in a matter of minutes. There is no need to wait days for an order to be processed and then to always be concerned that it will not arrive in time either for a paper deadline or class assignment. It would eliminate the waiting that takes place when the user discovers that another data collection would have been a better choice than the one originally requested.

It certainly is true that if public data networks worked in practice as they sound in theory, our data exchange problems would be over. However, some of the same problems that impact other media are also at work here. The speed with which data arrive over the lines is the result of a number of factors, including the different routes they must pass through to get from the source to the user's machine. The speed with which the data make that journey will be only as fast as the slowest link along the network path. Therefore, users never actually experience the maximum data transmittal speeds quoted for any given network. While every effort possible is made by the public data networks to assure complete transfer of data sent, transmittal problems can arise, resulting in incomplete transfers. The machine receiving the data must have space to accommodate the information coming down the lines. Finally, it is likely that not all data formats will readily lend themselves to public data networks. For example, since most of the data going over the lines are ASCII, EBCDIC binary data such as that found in OSIRIS dictionary files and other similar formats will certainly not be usable on the receiving end.

After looking at each of the several media available and the advantages and disadvantages of each, one may very well ask which is the best approach. We at the ICPSR have been spending a fair amount of time exploring each of the different formats. Unfortunately, we have not found any simple solution that will provide everything for everyone. Instead, we have concluded that we will be fortunate if we can provide something for everyone.

For the foreseeable future, ICPSR expects that much of the data supplied to users will continue to be provided on magnetic tape. All surveys of our users indicate that magnetic tape remains the overwhelming preference as a transmittal medium by the majority of our users. (However, there is a great deal of effort going into determining the next generation of reliable storage and transmittal media.) Additionally a significant number of our collections will not be suitable for transmittal by any other medium for the foreseeable future. This is largely due to the size of many of the collections which span several reels of tape. It is expected that the format of some collections will initially preclude their transmittal on other media. Hierarchical data files will be best supplied on magnetic tape at least for the time being. Nevertheless, the ICPSR has been taking steps to move toward other alternate media.

In February, 1991 nearly 100 copies of a two-volume CD-ROM containing the Panel Study of Income Dynamics data for waves 1-20 were distributed to Official Representatives at member institutions who had expressed a wish to participate in a field test of the product. The data were

supplied on CD-ROM in raw, ASCII format. SPSS/PC and SAS/PC statements were provided on each CD-ROM. Users could use the statements to prepare extracts from the main file or they were free to utilize their own software to perform the extracts. Responses on the questionnaire that was provided with the field-test copies indicated that users overwhelming approved of this approach for the CD-ROM.

ICPSR expects to produce a limited number of CD-ROMs in the future. It is expected that collections selected for this medium will be those that users have indicated they would like to see in this format, and those collections that have a high distribution volume. Some of these additional CD-ROM products should be available within a year.

The Computer Support Group of the ICPSR is in the process of conducting tests with a group of sites to evaluate the transmittal of ICPSR data through Internet. When these tests are concluded and relevant programming that supports this activity completed, we expect that access to ICPSR data will be expanded to include public data networks. It is expected that while eventually all ICPSR data will be accessible through the public data network, initially selected collections will be available in this manner. In order to make ICPSR data available through Internet, the thousands of files in our holdings will need to be moved from exclusively magnetic tape storage to optical disk storage. Since this will be a relatively large task, not all data will be stored on optical media immediately.

Finally, ICPSR is in the process of identifying collections that will additionally be available to users on floppy diskette. Data collections selected for floppy diskette production will be those for which there is demand for the data to be provided on this medium. The data collections that will be provided on diskette will be those that do not require large numbers of diskettes to accommodate a given data file. Additionally they will be collections that are available in raw, ASCII format. It is expected that the data will be supplied with self-extracting compression software with the appropriate README files that provide users with relevant information about the contents of the diskette.

As ICPSR continues with the installation of both software and hardware that will allow us to move from magnetic to optical storage media, we expect to continue to explore the feasibility of adding new services and upgrading older ones. While the ICPSR will monitor and respond to the technical changes many of our users are experiencing, we will also continue to remain responsive to those users who are not experiencing rapid technological changes. For the foreseeable future, ICPSR will seek a balance that allows us to serve users spanning the full spectrum of technical capabilities.

[1] Presented at the IASSIST 91 Conference held in Edmonton, Alberta, Canada. May 14 - 17, 1991.

An Analysis Of Cd-ROM as a Long Term Archiving Solution

by Denis Oudard[1]
Digipress

Before we go into this analysis, let's take a look at the history of archiving. This table shows different systems of communication and the role of each element as well as their similarities.

Data	Medium	Coding
Hieroglyph	**Stone/Papyrus**	**Rosetta Stone**
Text and B/W Image	**Microforms**	**Language, Alphabet, Magnifying Glass**
Databases, Raster Images, ASCII	**9 Track Tapes**	**Tape Drive, Computer System**
Sound, Digital and	**Compact Disc**	**CD-Player**

All of this combined and/or further processed provides us with information. There is no doubt that at the eve of the information age, we will be called upon to archive, for the long term, huge amounts of information.

In my first paper on this subject I stated that to provide a long term archiving solution, one needs to have two elements:

1. A retrieval or access system that will also endure the test of time.

2. A long lasting medium on which the data is stored.

One without the other is as useless as a deck of cards without aces.

Let's first tackle the 1st objective: access system longevity. Simplicity is the best way to insure the longevity of an access system. Such is the case for microforms. The magnifying glass is a simple system and we can rest assured that humanity will know how to use a magnifying glass for years to come. Therefore, the survival of this access system is virtually guaranteed.

Unfortunately, when dealing with computer archives, simplicity is definitely not part of the equation, so we have to look elsewhere for longevity factors.

The first one is momentum. In other words "How much acceptance or wide spread use does the technology have?" The law of large numbers is going to be a key ingredient. Consumer products have more momentum than professional products because they are sold in the 100s of millions rather than in the tens or hundreds of thousands. CD-ROM is the first computer media to be based on a widespread consumer item. Therefore, we have 100s of millions of machines out there that contain 90% of the key components of a CD-ROM drive. Even if CD-ROM technology, as we know it today, is abandoned in twenty

years or so, chances are we will find working CD-ROM drives in a 100 years.

Another important way momentum can be measured is by the number of manufacturers which build the same product. Today I can give you the names of at least 10 manufacturers of ISO 9660 CD-ROM drives. Are you ready? Here we go: Chinon, Digital Equipment, Hewlett Packard, Hitachi, NEC, Philips (LSMI), Pioneer, Sony, Texel, Toshiba; and I know I have left out some.

Compared to the momentum behind CD-ROM, other media are very fragile. WORM drives, for example, depend on the whims of a single manufacturer. Each manufacturer makes a different type of WORM, and the day the manufacturer stops making that drive, you have less than five years to transfer or lose your data.

The second critical factor relating to long-term access is system independence. The reason this is important is because none of the computer systems we know today, NONE, will be available in 20 years, let alone 50, 100 or 200.

In today's computer world we are dealing primarily with monolithic systems. That is, the hardware, operating system, application, and data are interdependent. IBM or VAX or most any other computer hardware have their own operating systems, their own application, and their own data sets, which can work only within their own world. The perfect archive needs to be accessible not only to all the systems that exist today, but also to all the systems which have yet to be created, all the super fast computers of tomorrow.

Thanks to the Red Book Standard and IS0 9660, CD-ROM is a peripheral device that has already achieved hardware independence. Diskettes, one of the most widely used media in the computer world, will never be able to make this claim. Try putting a MS/DOS diskette into a Macintosh; you can't even get a directory listing, let alone read the files. The 9 track tape might be the only other medium which has achieved the same degree of hardware independence.

Hardware independence is key. An ISO 9660 drive will read any ISO 9660 disk, regardless of the drive manufacturer and machine environment. That is an amazing feat, and again only equalled by the 9 track tape. That is the good news. The bad news is that while hardware independence is a big hurdle, it is only the first hurdle. True system independence demands much more. System independence in today's world means flexible interaction between the following elements:

1. Presentation or Process Software.
2. Retrieval Software.
3. The Information Itself.

This independence can only be achieved by setting standards which define how one element works with another. In essence, this is what the Red Book Standard does at the physical and opto-electronic levels. For logical software transactions, a similar set of standards is needed. Some of these standards have been established, and still more are emerging at this time.

I will limit my discussion to these three standards:

CD-RDx, SGML and TIFF. More should most likely be included, but these will serve the purpose of explaining how such standards work.

Let's start with CD-RDx. This standard was written for the primary purpose of enabling the user to utilize a single interface, his own. To achieve this, CD-RDx uses a client/server approach. The client (the user interface software, or application software) is separated from the server (also known as the retrieval engine). The idea is beautifully simple. If you establish a set of rules by which the client can ask the server for specific "searches" (e.g. a Boolean search), then any client using this protocol can interact with any server using the same protocol. Moreover, CD-RDx is set up in such a manner that the client and the server do not need to be on the same computer, not even on two computers with the same operating system. So now we have system independence between the retrieval engine system and the presentation or application system.

This represents important progress, but system independent data has not yet been achieved. The goal, remember, is to access the data on any disc via any retrieval engine. The solution is to agree on the structuring of information. TIFF and SGML are such standards which could be used to structure information in a universal, non-proprietary way.

Then, a CD-RDx compliant search engine could be developed to access any set of information structured within the guidelines of SGML and TIFF.

There it is, not simple, but resilient. For this to truly work, one also needs to take into account the problem of indexes, though the same reasoning applies, the question of standard index methodology is a truly thorny issue. What we would have is a structure where when one system changes, the information does not need to change and can remain on the same medium.

If the user/client system changes, client software is rewritten to run on the new system, along the guidelines of CD-RDx. If the server system changes, CD-RDx server

software is rewritten to retrieve data structured along the SGML and TIFF guidelines, keeping the data unchanged.

This arrangement is being developed today for easy distribution of data. The multitude of clients is due to the multitude of users and potential users of the data distributed, multiplied by the number of titles they each use. In the case of long term archiving, the multitude of systems is even greater because time is a new multiplying factor.

It must also be noted that this solution enables computer user interface and search engine to progress at their own pace, while "stable data sets" stay on one medium. Of course, the characteristics of the medium remain intact. The time will come though, when today's advanced CD-ROM technology will be regarded as a bulky and slow medium. Isn't it a lesser evil though, next to the alternative of losing access to the data content forever?

While today, CD-ROM is the perfect tool to have information on-line and on-site, to be used as an archive, the CD-ROM will be off-line. These archives will be fed into the super computers of tomorrow. After all, the CD-ROM will not always be the medium where the information resides for processing. Often the needed data will be downloaded to fast access, massive storage devices of future computers for processing as needed. This already happens when information is downloaded from a CD-ROM onto a word processor or a spreadsheet.

To recap the longevity factors relating to access systems, we have established that.

Hardware availability in the distant future is all the more likely if the technology is:

1. Widely Accepted- thanks to the support of a product in the consumer market that uses the same technology. Proliferation helping the survival of functioning hardware.

2. The Technology Must Also Be Standardized- which helps the consumer market become even bigger (see 1.) and documents the detailed working of hardware in a precise and widely available manner.

Logical access in the distant future is all the more likely if:

1. Information is system independent.

2. Information is formatted using a non-proprietary standard. Indeed the wide acceptance of a few well chosen standards will foster the development of compatible software and ensure consistent data structuring.

3. The users can have access to this data and then manage it (for display, printing or processing) using the system of their choice through a system independent client/server type of protocol.

More to the point, we have established that CD-ROM is the medium that answers these requirements. The best 9 track tape, while very widely used, does not benefit from the momentum of a consumer market. WORM does not even come into the picture. This is not to say though that WORM, tape and other media, which do not meet some or any of these criteria, will not continue to perform important tasks in our computer rooms.

It is now time to move onto the longevity of the media itself. Of course, given the above conclusion, if one could find a CD-ROM that would last a hundred years or more, we would be home free. As some of you in the audience know, I am with a company that has dedicated large amounts of time and money over the last 5 years to develop such a CD-ROM.

Conclusion

As I see it, CD-ROM, along with fantastic information distribution capabilities, has already more long term archiving features than any other mass computer storage available. Only a few additional steps need to be taken to truly make it one of the top answers to data information archiving for the next 20 years. These steps are the focus of a committee being considered for creation by the Commission on Preservation and Access. All this is very encouraging, and I hope that it can help solve some of your long term archiving needs.

Bibliography:

1

"Taking a Byte out of History: The Archival Preservation of Federal Computer Records", House Report 101-978, 101st Congress, 2nd Session

"GAO Faults NASA for Mismanaging Storage of Valuable U.S. Space Science Data" James R. Asker, Aviation Week & Space Technology, April 2, 1990 (enclosed)

"Lost in Tapes" J. Sniffen, Associated Press, January 2, 1991 (enclosed)

"National Archives Needs Better Record-Keeping Technology, Report Says" Ann M. Mercier, Federal Computer Weekly, November 1990

"'SGML Like'- And We Could Get 'Sort of Married' Too.." William Zoellick, Disc magazine, Premier Issue, Fall 1990. page 53-54. Available from Helgerson Associates. Tel: (703) 237-0682

6"CD-ROM Read-Only Data Exchange Standard, Version 3.0"; December 30, 1990; available from the OPA. Tel: (614) 793-9660

"An Analysis of Compact Discs As a Long term Archiving Solution" Denis Oudard, American Library Association Midwinter Meeting, ALCTS-PLMS' Physical Quality & Treatment Discussion Group, January 12, 1991

"CD-ROM as an Archiving Medium?" Denis Oudard, Working paper, Digipress, February 1991. Available on demand. Tel: (502) 895-0565Presented at the IASSIST 91 Conference held in Edmonton, Alberta, Canada. May 14 - 17, 1991.

[1] Presented at the IASSIST 91 Conference held in Edmonton, Alberta, Canada. May 14 - 17, 1991. For further information on Digipress's Century Disc contack the author at: 2016 Bainbridge Row Drive, Louisville, Kentucky 40207 USA. Tel (502) 895-0565.

Electronic Reference Systems in the Year 2000: The Symposium on Advanced Information Processing & Analysis, March 24 - 26, 1992

by Lee A. Gladwin[1], Center for Electronic Records (NNX)
National Archives and Records Administration, Washington, DC 20408

"I'm drowning in open source information!" is the cry of intelligence analysts, a cry not unfamiliar to many other researchers. How does one navigate through a turbulent paper sea? Confronted with a wealth of new open source material (e.g. newspapers, television, technical journals, wire service bulletins, etc.), analysts, who should be interpreting incoming information, currently spend most of their time either sifting and reading documents or writing reports based upon them. Contributing to their problems are those of inefficient technological transfer, shrinking resources (funds and people), overlapping R&D efforts, and the lack of system integration. Unlike most researchers, however, intelligence community has an organization, the Advanced Information Processing & Analysis Steering Group (AIPASG), through which to present its needs to potential contractors and enough funding to attract bidders. AIPASG held its annual meeting March 24 - 26, 1992 in Reston, Virginia.

As with other researchers, analysts need assistance in scanning the data, selecting and organizing documents relevant to their problem, and printing the results. What they do not need is to spend time battling a recalcitrant retrieval system, reading system manuals, attending software workshops, or talking with customer service about software problems.

Through panel presentations, contractors were acquainted with technical developments in the five areas of need identified by AIPASG:

- Intuitive user interfaces
- Document processing, organization and management
- Transparent access to multiple data bases and sources
- Collaborative communications
- Automated data understanding

The first need is for a powerful but easily used interface which does not require exhaustive efforts to accomplish simple procedures. Any system should be developed in close consultation with the users, not in a vacuum.

Document processing addresses the need to organize, sort and link documents relevant to an intelligence problem before routing them to the analyst studying those problems. Central to the solution of the problem of coping with massive amounts of material is to shift the focus from document retrieval to managing discrete pieces of information, using visual indicators to point to other possibly relevant sources.

Transparent access to multiple databases addresses the need to know what data is out there and how to retrieve it.

Collaborative communications concerns the institutional barriers to sharing information; i.e., security, organizational territoriality, etc.

Symposium topics addressed these five needs and such key technologies as continuous speech recognition, natural language and graphical user interfaces, on-line tutors with user modelling capabilities, optical character recognition, automated data extraction, and multiple database correlation. Some of the technologies relating to this future system are described in the concluding portion of this report.

Natural Language/Text Processing

Papers presented in this session addressed the needs to translate text from foreign languages and then extract information and present it to the user in a meaningful manner. Programs were described for translating news items written in Spanish and Japanese, parsing the text syntactically and semantically, and displaying information in an on-screen template. Adrian Kleiboemer, MITRE, called for creation of reusable environments which could be ported easily to any number of applications without having to build a new natural language frontend (NLF) for every new application as is currently being done.

Natural language frontends are currently available for searching large databases without forcing the user to learn SQL. Natural Language Inc. developed a frontend for ORACLE which takes short phrases and even sen-

tences, parses them into SQL statements, and runs them against the database. They may be used in conjunction with graphical user interfaces (GUIs).

Optical Characters Recognition and Neural Networks
The CIA currently is engaged in a five - six year research program aimed at translating source documents, in varying conditions and all languages, into machine-readable format. There are two forms of OCR enhancement: Digital and repair. Digital enhancement clarifies the image using bit mapping and a gray scale. Since it simply prints what is there, letters may be broken or run together. Digital enhancement cannot recognize letters or words. Repair enhancement techniques seek to reconstruct letters and words from faded, damaged or crooked images (i.e., paper orientation). An IBM program was described which uses neural networks to segment pages into regions (e.g., return address, recipient address, stamp, logo, signature), identifies and classifies these segments, and deduces whether the document is a letter, form, article, etc. Neural networks (computer simulated biological neurons) are used in pattern recognition tasks to identify printed or written text images. Applications are currently underway at the US Post Office to read handwritten addresses.

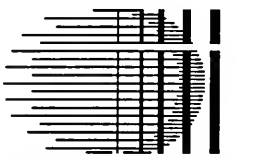
Document Processing, Organization & Management
Papers presented in this session dealt with the "derivation and use of statistical procedures for retrieval and data extraction". Richard M. Tong presented a paper entitled "Automatic Document Retrieval Using CART [Classification and Regression Tree]" which classifies and retrieves documents containing at least one of 15 specific sub-concepts of "civil unrest" e.g., "labor union". Their initial results show better retrieval results for concept-based search than by using standard key word searches in a test involving one-thousand news items. Relevant to this finding was a remark made by Paul Thompson, an attendee, who observed that Boolean or probabilistic ranking approaches to document retrieval are insufficient to assure a document's relevance. Simply adding terms to an SQL query only serves to increase errors.

Data Bases & Information Retrieval
Large scale information retrieval (LSIR) addresses the need to retrieve information from "data sources" that are complex and distributed globally or through different departments of an organization. Potential technologies for dealing with an "indexless encyclopedia" are object-oriented databases, hypermedia, natural language processing, parallel processing, expert systems, and information visualization.

In addition to meeting the five needs, it was suggested that intelligent user interfaces be developed which could model user expertise, and, in the event of error, infer what the researcher was attempting to do and provide greater assistance in information extraction; i.e., an electronic reference guide sensitive to the researcher's facial expressions and body language.

While this electronic reference guide is not yet fully integrated, many of the components are under development and were discussed at the symposium. Some day an electronic reference guide will help researchers navigate their ways through text, graphics, art, musical recordings, still and motion pictures in search of information relevant to the problem at hand. Undiscussed were questions about the implications of this technology for researchers, research methodology and the reference room in the year 2000. Perhaps now would be a good time to begin thinking about these implications.

[1] Article based on notes taken at The Symposium on Advanced Information Processing & Analysis, held in Reston, Virginia, March 24 - 26, 1992.



IASSIST

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICE AND TECHNOLOGY

• • • •

ASSOCIATION INTERNATIONALE POUR LES SERVICES ET TECHNIQUES D'INFORMATION EN SCIENCES SOCIALES

Membership form

The International Association for Social Science Information Services and Technology (**IASSIST**) is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the **IASSIST QUARTERLY**. They also benefit from reduced fees for attendance at regional and international conferences sponsored by **IASSIST**.

Membership fees are:
Regular Membership. $20.00 per calendar year.
Student Membership: $10.00 per calendar year.

Institutional subcriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subcription:
$35.00 per calendar year (includes one volume of the Quarterly)

I would like to become a member of IASSIST. Please see my choice below:

- ☐ $40 Regular Membership
- ☐ $20 Student Membership
- ☐ $70 Institutional Membership

My primary Interests are:

- ☐ Archive Services/Administration
- ☐ Data Processing
- ☐ Data Management
- ☐ Research Applications
- ☐ Other (specify)_____

Please make checks payable to IASSIST and Mail to :

Ms Kay Worrell
Treasurer, IASSIST
% The Conference Board
845 Third Avenue
New York, NY 10022-6601

Name / title

Institutional Affiliation

Mailing Address

City

Country / zip/ postal code / phone

BOOK RATE

BOOK RATE